

Characteristic Kernels and Infinitely Divisible Distributions

Yu Nishiyama

NISHIYAM@ISM.AC.JP

*Research Center for Statistical Machine Learning
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan*

Kenji Fukumizu

FUKUMIZU@ISM.AC.JP

*The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan*

Editor:

Abstract

This paper connects the shift invariant characteristic kernels on \mathbb{R}^d to the infinitely divisible distributions of probability theory. Recently, the embedding of probabilistic distributions into a reproducing kernel Hilbert space (RKHS) has been studied in machine learning, called the kernel mean. The characteristic kernel is such that it maps any two different probability distributions to different elements in the RKHS, i.e., the mapping is injective. This property is essential to a variety of kernel mean applications, such as hypothesis tests, Bayesian inference, classification, dimension reduction, and reinforcement learning.

In this paper, we show that in general the shift invariant kernel on \mathbb{R}^d is characteristic if the kernel is generated by the bounded continuous density of a symmetric and infinitely divisible distribution. This class includes kernels of the Gaussian distribution, Laplace distribution, α -stable distribution ($0 < \alpha < 2$), Student's t -distribution, and generalized hyperbolic (GH) distribution. We call this class the *convolution infinitely divisible (CID)* kernel.

Under the guarantee that the CID kernel is characteristic, we consider kernel means of probabilistic models of infinitely divisible distributions. This includes the Gaussian distribution, α -stable distribution, NIG distribution, and VG distribution. The kernel mean and relevant inner products have the same density form with different parameter values when the RKHS is chosen as a corresponding RKHS. The simple density form makes the computation of inner products feasible. This provides a notion of the conjugate kernel to probabilistic models in the sense of the kernel mean.

Keywords: Characteristic Kernel, Kernel Mean, Infinitely Divisible Distribution, Stable Distribution, GH Distribution

1. Introduction

The kernel method has provided many nonlinear algorithms in machine learning by manipulating data in the reproducing kernel Hilbert space (RKHS) (Schölkopf and Smola, 2002; Steinwart and Christmann, 2008). In the kernel method, the data are mapped to elements in the RKHS by a positive-definite (p.d.) kernel. Recently, the embedding of a probability distribution in an RKHS, called the *kernel mean* (Fukumizu et al., 2013), has been applied to various problems, such as density estimations (Smola et al., 2007; Song et al., 2008; McCalman et al., 2013), hypothesis tests (two sample test, Gretton et al. 2012, indepen-

dence test, Gretton and Györfi 2010, conditional independence test, Fukumizu et al. 2008), Bayesian inference (conditional distributions, Song et al. 2009, hidden Markov models, Song et al. 2010, belief propagation, Song et al. 2011, Bayes' rule, Fukumizu et al. 2013), classification (Muandet et al., 2012), dimension reduction (Fukumizu and Leng, 2012), and control problems (Markov decision process, Grünewälder et al. 2012, partially observable Markov decision process, Nishiyama et al. 2012, path integral control, Rawlik et al. 2013, and predictive state representation, Boots et al. 2013). There is a good review paper on Bayesian inference with the kernel mean (Song et al., 2013).

Let $\mathcal{P}(\mathcal{X})$ be the set of probability distributions on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Let X denote a random variable taking values in \mathcal{X} with law $P \in \mathcal{P}(\mathcal{X})$. The kernel mean m_P of a probability distribution P in an RKHS \mathcal{H} associated with a p.d. kernel $k_{\mathcal{X}}(\cdot, \cdot)$ on a set \mathcal{X} , is defined by the mean element $m_P := \mathbb{E}_X[k_{\mathcal{X}}(\cdot, X)]$ in the RKHS \mathcal{H} . The kernel mean m_P determines a mapping $P \mapsto m_P$ from the set $\mathcal{P}(\mathcal{X})$ to the RKHS \mathcal{H} . The mapping is called the *kernel mean map*.

A desirable property for representing probability distributions $P \in \mathcal{P}(\mathcal{X})$ in an RKHS \mathcal{H} is that different probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$, $P \neq Q$ map to different elements $m_P \neq m_Q$ in the RKHS \mathcal{H} , i.e., the kernel mean map $P \mapsto m_P$ is injective. If two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$ imply the same RKHS element $m_P = m_Q$, we may easily encounter problems. In the kernel two-sample test (Gretton et al., 2012), two datasets are tested to determine whether they are drawn from the homogeneous probability distribution, using the difference $\|m_P - m_Q\|_{\mathcal{H}_{\mathcal{X}}}$ of the two RKHS representations, i.e., the norm of the two kernel means. If the two probability distributions imply the same kernel mean, they are not distinguishable. The injective kernel mean map is important because it endows a metric in the set $\mathcal{P}(\mathcal{X})$ of probability measures via the Hilbert space norm $\|\cdot\|_{\mathcal{H}_{\mathcal{X}}}$, i.e., the Hilbertian metric (Hein and Bousquet, 2005; Sriperumbudur et al., 2010). This metric is studied with comparing it to other popular metrics in the set of probability measures (Sriperumbudur et al., 2010).

The injectivity of the kernel mean map $P \mapsto m_P$ depends on the choice of the p.d. kernel $k_{\mathcal{X}}(\cdot, \cdot)$ as $m_P = \mathbb{E}_X[k_{\mathcal{X}}(\cdot, X)]$. A p.d. kernel $k_{\mathcal{X}}(\cdot, \cdot)$ is called *characteristic* if the kernel mean map is injective (Fukumizu et al., 2008). Basically, all of the above applications assume the use of characteristic kernels. The RKHS \mathcal{H} generated by a characteristic kernel is called the *characteristic RKHS*, which is a sufficiently rich space to distinguish all probability distributions $\mathcal{P}(\mathcal{X})$ in the RKHS \mathcal{H} . It is known that the RKHS \mathcal{H} on a general measurable space $(\mathcal{X}, \mathcal{B})$ is characteristic if and only if $\mathcal{H} + \mathbb{R}$ (the direct sum of the two RKHSs) is dense in the q -integrable function space $L^q(\mathcal{X}, P)$ for every probability measure P on $(\mathcal{X}, \mathcal{B})$ with some $q \geq 1$ (Fukumizu et al., 2008, 2009a, Lemma 1, Proposition 5, respectively),

When is a p.d. kernel $k_{\mathcal{X}}(\cdot, \cdot)$ characteristic? The conditions are explored according to the type of domain \mathcal{X} and the class of kernel (Gretton et al., 2007; Fukumizu et al., 2008, 2009a,b; Sriperumbudur et al., 2010). An overview of this is given in (Sriperumbudur et al., 2010). In this paper, we focus on the Euclidean space $\mathcal{X} = \mathbb{R}^d$, and consider the class of *shift-invariant* kernels. The class includes typical examples of the Gaussian and Laplace kernels on \mathbb{R}^d , and is one of the important classes. In this class, a necessary and sufficient condition for a p.d. kernel $k_{\mathcal{X}}(\cdot, \cdot)$ to be characteristic is known in terms of the description

of the Bochner theorem. This states that a bounded continuous shift-invariant kernel on \mathbb{R}^d is characteristic if and only if the support of the Fourier transform Λ of the positive-definite function ψ is the entire \mathbb{R}^d (Sriperumbudur et al., 2010).

In this paper, we relate the class of shift-invariant characteristic kernels on \mathbb{R}^d to the infinitely divisible distributions. A probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$ is called *infinitely divisible* if, for every positive integer $n > 0$, there is a probability distribution $P_n \in \mathcal{P}(\mathbb{R}^d)$ such that P can be expressed as the n -fold convolution of P_n . Infinitely divisible distributions are important in limit theorem. According to the *Lévy-Khintchine formula*, every infinitely divisible distribution P on \mathbb{R}^d can be represented by the *generating triplet* (A, ν, γ) , where A is a symmetric nonnegative-definite $d \times d$ matrix, $\gamma \in \mathbb{R}^d$ is a vector, and ν is a *Lévy measure* (Sato, 1999). As given in the contribution Section 1.2, we show that in general the p.d. kernel associated with a bounded continuous density of a symmetric infinitely divisible distribution on \mathbb{R}^d is characteristic. See Section 1.2 for this.

Next, we consider kernel means of probabilistic models¹. In the various machine learning applications using kernel means presented above, it is important to compute the following inner products w.r.t. kernel means m_P and m_Q :

$$\langle m_P, k(\cdot, x) \rangle_{\mathcal{H}}, \quad \langle m_P, m_Q \rangle_{\mathcal{H}}, \quad P, Q \in \mathcal{P}(\mathbb{R}^d). \quad (1)$$

Typically, one estimates kernel mean m_P using a linear combination form $\hat{m}_P = \sum_{i=1}^n w_i k_{\mathcal{X}}(\cdot, X_i)$ of feature maps $\{k_{\mathcal{X}}(\cdot, X_i)\}_{i=1}^n$ with data $\{X_i\}_{i=1}^n$. This allows us to efficiently compute (1), with avoiding the intractable computation of high-dimensional integrals of (1), by virtue of the reproducing property of the RKHS \mathcal{H} or so-called kernel trick in the kernel method literature. The computation of inner products (1) then results in computing kernel $k_{\mathcal{X}}(X_i, X_j)$ and the Gram matrices $G = (k_{\mathcal{X}}(X_i, X_j))_{ij}$. Note that if inner products (1) can be efficiently computed, other functions such as the norm

$$\|m_P - m_Q\|_{\mathcal{H}}^2 = \|m_P\|_{\mathcal{H}}^2 + \|m_Q\|_{\mathcal{H}}^2 - 2\langle m_P, m_Q \rangle_{\mathcal{H}}, \quad (2)$$

can also be efficiently computed.

On the other hand, it is also the case that the kernel mean m_P is not represented as a linear combination of feature maps as above. The kernel mean of a probabilistic model is a case. The kernel mean m_P of a Gaussian distribution P in the Gaussian RKHS \mathcal{H} generated by a Gaussian kernel $k(\cdot, \cdot)$ is also a Gaussian, and is not the form of the linear combination of feature maps. In the Gaussian case, inner products (1) can be expressed as Gaussians and the computation is feasible. Beyond the Gaussian case, it will be useful to show a list how to compute the basic inner products (1) and (2) when P are various probabilistic models, so that more flexible computations of kernel means with probabilistic models are possible.

Specifically, kernel means m_P of probabilistic models P have been used in the following applications. Smola et al. (2007), Song et al. (2008), and McCalman et al. (2013) consider the recovery of the density $\hat{p}(x)$ from a given (estimated) kernel mean m_P in an RKHS \mathcal{H} . The density $p(x)$ is typically recovered using a mixture of base probabilistic models, such as a mixture of Gaussians. The task requires the computation of kernel means

1. In particular, this paper focuses on probabilistic models of infinitely divisible distributions.

of base probabilistic models. The support measure machine (SMM) (Smola et al., 2007; Muandet et al., 2012) also uses the kernel means of probabilistic models. These detailed tasks are described in Section 1.1. Besides these applications, kernel means of probabilistic models will be ubiquitous in various applications where the kernel mean approach and the probabilistic model approach are compatibly combined. An important application will be the development of the semi-parametric kernel mean Bayesian inference, based on the currently developed full-nonparametric kernel mean Bayesian inference (Song et al., 2009, 2010, 2011; Fukumizu et al., 2013), by incorporating probabilistic models into them.

Under the above background and motivations, the contribution of this paper is given in Section 1.2 and Section 1.3.

1.1 Motivation Examples: Kernel Mean of Probabilistic Model

The inner products (1) of probabilistic models $P, Q \in \mathcal{P}(\mathcal{X})$ are used in the following tasks.

Density Estimation: (Smola et al., 2007; Song et al., 2008; McCalman et al., 2013)

Let $m[P]$ and $m[p]$ denote the kernel mean of a probability distribution P and density p in an RKHS \mathcal{H} , respectively. The kernel mean $m[P]$ is often estimated from data such as the conditional kernel mean estimator (Song et al., 2009; Fukumizu et al., 2011) or the posterior kernel mean estimator (Fukumizu et al., 2011) in, e.g., kernel Bayesian inference. Given a kernel mean estimator $\hat{m}[P]$, the task is to recover the underlying probability distribution P . One approach to the estimation of P is to use a mixture of probabilistic models, e.g., a mixture of Gaussians. Let $\hat{m}[P]$ be a given estimator, and p be a convex combination of a set of M candidate probabilistic model densities $\{p_i\}$, i.e.,

$$p = \sum_{i=1}^M \alpha_i p_i, \quad \alpha^\top \mathbf{1} = 1, \quad \alpha \geq 0, \quad (3)$$

where $\mathbf{1} \in \mathbb{R}^M$ is the vector composed entirely of 1s, and $\alpha \geq 0$ denotes element-wise inequality. Under the fixed M candidate densities $\{p_i\}$, p is estimated by solving the following optimization problem for weights α :

$$\min_{\alpha} \|\hat{m}[P] - m[p]\|_{\mathcal{H}}^2 + \Omega[\alpha], \quad \text{subject to } \alpha^\top \mathbf{1} = 1, \alpha \geq 0, \quad (4)$$

where $\Omega[\alpha]$ is a regularizer to control the smoothness of the minimizer, such as $\frac{\lambda}{2} \|\alpha\|^2$ ($\lambda > 0$) (the candidate densities were also attempted to be optimized in Song et al. 2008.). It is known (Smola et al., 2007) that the optimization (4) is a quadratic program, and reduces to

$$\min_{\alpha} \frac{1}{2} \alpha^\top (Q + \lambda I) \alpha - l^\top \alpha \quad \text{subject to } \alpha^\top \mathbf{1} = 1, \alpha \geq 0, \quad (5)$$

where I is the $M \times M$ identity matrix, and the vector $l \in \mathbb{R}^M$ and matrix $Q \in \mathbb{R}^{M \times M}$ are given by

$$\begin{aligned} l_j &= \langle \hat{m}[P], m[p_j] \rangle_{\mathcal{H}}, & j &\in \{1, \dots, M\}, \\ Q_{ij} &= \langle m[p_i], m[p_j] \rangle_{\mathcal{H}}, & i, j &\in \{1, \dots, M\}. \end{aligned} \quad (6)$$

The given kernel mean estimator $\hat{m}[P]$ often has a weighted sum form $\hat{m}[P] = \sum_{i=1}^n w_i k(\cdot, X_i)$, with data $\{X_i\}$ and weight vector $w \in \mathbb{R}^n$ (Song et al., 2009; Fukumizu et al., 2011). Then, the vector l reduces to

$$l_j = \langle \hat{m}[P], m[p_j] \rangle_{\mathcal{H}} = \sum_{i=1}^n w_i \langle k(x_i, \cdot), m[p_j] \rangle_{\mathcal{H}}, \quad j \in \{1, \dots, M\}. \quad (7)$$

Inner products of kernel means of probabilistic models, i.e., (1), appear in (6) and (7). In Song et al. (2008) and McCalman et al. (2013), Gaussian densities are used for the candidate densities p_i , thus producing the closed-form solution of the kernel mean.

There can be other options for the candidate densities, such as using heavy tailed or semi-heavy tailed densities of α -stable densities and zero-skewed NIG (VG) densities, considered in this paper. If the true underlying distribution P in $\hat{m}[P]$ is (semi-)heavy tailed, then recovering the density using the mixture of Gaussians will not provide a sufficiently good density estimation. Recovering densities of (semi-)heavy tailed distributions P can be approached.

Support Measure Machine (SMM): (Smola et al., 2007; Muandet et al., 2012)

The SMM (Muandet et al., 2012) is a supervised learning algorithm whose input is a probability distribution, rather than the finite dimensional vector in SVMs. Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on the nonempty set \mathcal{X} . Let \mathcal{Y} be the set of labels, typically the binary classification $\mathcal{Y} = \{+1, -1\}$. A datum consists of an input probability distribution $P \in \mathcal{P}(\mathcal{X})$ and its output $y \in \mathcal{Y}$, i.e., a pair (P, y) . Given a set of data $(P_i, y_i)_{i=1}^m$, SMM learns a function $h : \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{Y}$. It is necessary for the SMM to set a kernel $K : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$ on the probability measures to measure the similarity among two probability distributions. Muandet et al. (2012) used the Hilbertian metric $K(P, Q) = \langle m_P, m_Q \rangle_{\mathcal{H}}$, $P, Q \in \mathcal{P}(\mathcal{X})$ defined via the kernel mean m_P and m_Q , and considered more general kernels such as

$$K(P, Q) = (\langle m_P, m_Q \rangle_{\mathcal{H}} + c)^d, \quad (8)$$

$$K(P, Q) = \exp\left(-\frac{\gamma}{2} \|m_P - m_Q\|_{\mathcal{H}}^2\right). \quad (9)$$

Many standard nonlinear kernels can be considered as kernels on probability measures via $\langle m_P, m_Q \rangle_{\mathcal{H}}$. Muandet et al. (2012) refers to the kernel k that defines the kernel mean m_P as the *embedding kernel*, and to the final K as the *level-2 kernel*. When P and Q are Gaussians, the paper used the closed-form solution for the inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ as given in Section 4.1.

The inner product can also be computed in other cases when P and Q are (semi-)heavy tailed distributions of α -stable densities or zero-skewed NIG (VG) densities, using conjugate results considered in this paper.

1.2 Convolution Infinitely Divisible (CID) Kernel

The Gaussian and Laplace distributions are infinitely divisible, and the associated Gaussian and Laplace kernels are known to be characteristic. Is it generally true that the p.d. kernel associated with an infinitely divisible distribution is characteristic? The answer is

yes. We show that the p.d. kernel associated with a bounded continuous density of a symmetric infinitely divisible distribution on \mathbb{R}^d is characteristic (Theorem 3.3). The kernel examples are shown in Example 3.4. This class includes the symmetric α -stable (SaS) distributions ($0 < \alpha \leq 2$, $\alpha = 2$ corresponds to Gaussian distributions, $\alpha = 1$ corresponds to Cauchy distributions), symmetric Student's t -distributions, symmetric generalized hyperbolic (SGH) distributions, and symmetric tempered α -stable ($ST\alpha S$) distributions. These are well-known infinitely divisible distributions. The proof of the statement is immediate after applying the necessary and sufficient condition in terms of the Bochner theorem (Sriperumbudur et al., 2010) to the general Lévy–Khinchine formula.

We call this class the *convolution infinitely divisible* (CID) kernels. CID kernels are characteristic. We mention an important subclass of infinitely divisible distributions, called the *self-decomposable* distributions. An infinitely divisible distribution does not always have a density function, but a nondegenerate self-decomposable distribution is guaranteed to have a density function. We have a statement that the p.d. kernel associated with a continuous bounded nondegenerate self-decomposable density is characteristic (Corollary 3.5). We also mention how to generate a CID kernel using the symmetrization operation of an (asymmetric) infinitely divisible distribution (Section 3.3).

1.3 Kernel Mean of Probabilistic Model

Under the guarantee that the CID kernels are characteristic (Section 1.2), we consider kernel means of probabilistic models of infinitely divisible distributions. The kernel mean of a Gaussian distribution in a Gaussian RKHS is simply given by a Gaussian density form. This closed property originates from the convolution stability property of the Gaussian distribution. We first show that, from the general perspective, the kernel mean m_P of an infinitely divisible distribution P in the RKHS \mathcal{H} generated by a CID kernel is an infinitely divisible density (Proposition 3.6). Then, the kernel means and inner products (1) are described by the generating triplet (A, ν, γ) of the infinitely divisible distribution (Proposition 3.8). This expression is very general for computational purposes and is a conceptual understanding. Imposing structures on the generating triplet (A, ν, γ) produces various subclasses of infinitely divisible distributions. We focus on the following specific infinitely divisible distributions (Section 4):

- α -stable distributions on \mathbb{R}^d ($0 < \alpha \leq 2$; Gaussian ($\alpha = 2$), Cauchy ($\alpha = 1$)),
 - Sub-Gaussian α -stable distributions ($0 < \alpha < 2$)
 - Isotropic α -stable distributions ($0 < \alpha < 2$)
- Multivariate GH distributions on \mathbb{R}^d
 - Normal inverse Gaussian (NIG) distributions
 - Variance gamma (VG) distributions

Stable distributions are well-known infinitely divisible distributions². The GH, NIG, and VG distributions are well-known in mathematical finance and risk management (Schoutens, 2003; Cont and Tankov, 2004; Barndorff-Nielsen and Halgreen, 1990; Madan et al., 1998; Barndorff-Nielsen, 1998; Barndorff-Nielsen and Prause, 2001). The convolution stability property is known in these probabilistic models.

We show that, for each $0 < \alpha \leq 2$, the kernel mean m_P is given by an α -stable density when P is an α -stable distribution and the kernel $k(\cdot, \cdot)$ is an α -stable kernel (α -stable RKHS) (Propositions 4.10, 4.17). Then, we show that the computation of each inner product (1) is a computation of an α -stable density value (Propositions 4.12, 4.19). In the special case of the sub-Gaussian α -stable distributions, the closed property holds if probability distribution P and kernel $k(\cdot, \cdot)$ possesses the same associated positive definite matrix R up to a scalar multiplication $\sigma > 0$ (Propositions 4.22, 4.24).

Similarly, we show that the kernel mean m_P is given by a zero-skewed NIG (resp. VG) density when P is a zero-skewed NIG (resp. VG) distribution and the kernel $k(\cdot, \cdot)$ is a zero-skewed NIG (resp. VG) kernel (Proposition 4.30). The computation of each inner product (1) is a computation of a zero-skewed NIG (resp. VG) density value (Propositions A.4, A.5). Note that these hold only for the zero-skewed case. The Matérn kernel is a well-known kernel commonly used in spatial statistics and geostatistics. The Matérn kernel is a special case of the VG kernel.

Following the above observation, for convenience, we introduce the notion of the *conjugate* kernel to a probabilistic model in the sense of the kernel mean (Section 5). We say a p.d. kernel is conjugate to a probabilistic model if the kernel mean and the probabilistic model have the same density form with different parameter values³. We call the associated RKHS the conjugate RKHS to the probabilistic model. Similarly, we say a probabilistic model is conjugate to the kernel if the kernel is conjugate to the probabilistic model. The Gaussian kernel (Gaussian RKHS) is conjugate to the Gaussian distributions. For each $\alpha \in (0, 2]$, the α -stable kernel (α -stable RKHS) is conjugate to the α -stable distributions. Note that the stable characteristic exponent α must be identical for the α -stable distributions and the α -stable kernel. The RKHS generated by the NIG (resp. VG) kernel is conjugate to the zero-skewed NIG (resp. VG) distribution.

In the simulation side, as above, when P is an α -stable distribution and the kernel $k(\cdot, \cdot)$ is an α -stable kernel, the computation of each inner product (1) is a computation of an α -stable density value. Note that an α -stable density does not generally have the closed-form expression unlike Gaussian distributions, but the α -stable density value is numerically computable at least low dimensions or special cases. For example, the STABLE 5.1 software⁴ simulates the α -stable distribution to dimension $d = 2$ when P is the general

2. See John Nolan's Stable Distribution Page, <http://academic2.american.edu/~jpnolan/stable/stable.html>. Many of application areas and references of stable distributions can be found in <http://academic2.american.edu/~jpnolan/stable/StableBibliography.pdf>.

3. This is an analogy of the *conjugate prior* in the exponential family. In Bayes' theorem, the conjugate prior to the likelihood is such that the prior and the posterior have the same density form of probabilistic models, such as the Gaussian.

4. <http://academic2.american.edu/~jpnolan/stable/stable.html>. Detailed information is given in the user manual for STABLE 5.1 <http://www.robustanalysis.com/MatlabUserManual.pdf>.

α -stable case, and $d \leq 100$ when P is the isotropic or sub-Gaussian α -stable distribution case. The isotropic and sub-Gaussian α -stable densities are elliptically contoured and can be simply evaluated by the amplitude density of a scalar value. The R package, ghyp (Breyman and Lüthi, 2013), simulates the GH distribution. The zero-skewed GH density is also elliptical, and the computation of the density is simply via the univariate amplitude density. For further computational issue, see Section 6.

1.4 Organization of this Paper

The rest of this paper is organized as follows. The next section presents some notation and definitions for the kernel mean in the RKHS associated with a p.d. kernel and the characteristic kernel. Related facts are shown as preliminaries. In Section 3, we give the main result of this paper. We introduce the CID p.d. kernel, and show that it is characteristic. In Section 4, we see examples of the CID kernel, ranging from the simple Gaussian case on \mathbb{R}^d to the univariate α -stable distribution on \mathbb{R} , the multivariate α -stable distribution on \mathbb{R}^d (which includes the sub-Gaussian α -stable distribution and the isotropic α -stable distribution), and the rich GH distribution class. In Section 5, given several examples of kernel means that have the same density form as in Section 4, we introduce the notion of the conjugate kernel to probabilistic models in terms of the kernel mean. Section 6 mentions the computational issue for computing kernel means of probabilistic models. Finally, conclusions are presented in Section 7.

2. Kernel Mean in RKHS and Characteristic Kernel

This section introduces some preliminaries regarding p.d. kernels, reproducing kernel Hilbert spaces, kernel means, and characteristic kernels, and presents related results used in later sections. All functions considered in this paper are real-valued unless stated otherwise. We denote $\mathbb{R}_+ = [0, \infty)$.

Positive-definite (p.d.) kernel:

Let \mathcal{X} be an arbitrary nonempty set. A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is said to be symmetric if $k(x, y) = k(y, x)$ holds for any $x, y \in \mathcal{X}$. A symmetric function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a *positive-definite (p.d.) kernel* if, for any $n \in \mathbb{N}$ and any $x_1, \dots, x_n \in \mathcal{X}$, the $n \times n$ matrix G such that $G_{ij} = k(x_i, x_j)$, $i, j \in \{1, \dots, n\}$ is positive-semidefinite.

Reproducing kernel Hilbert space (RKHS):

A reproducing kernel Hilbert space (RKHS) \mathcal{H} on a nonempty set \mathcal{X} is a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ where, for every $x \in \mathcal{X}$, there exists a unique element $e_x \in \mathcal{H}$ such that

$$f(x) = \langle f, e_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H} \quad (\text{reproducing property}) \quad (10)$$

(equivalently, for every $x \in \mathcal{X}$, the linear functional $L_x : \mathcal{H} \rightarrow \mathbb{R}; f \mapsto f(x)$ is continuous). Here, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of the Hilbert space. For any RKHS \mathcal{H} , the function $k(x, y) = e_x(y)$, $x, y \in \mathcal{X}$, is a p.d. kernel. Conversely, for any p.d. kernel $k(x, y)$, $x, y \in \mathcal{X}$, there exists a unique RKHS \mathcal{H} [Moore–Aronszajn Theorem].

Kernel mean in RKHS on a measurable space:

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space, and $\mathcal{P}(\mathcal{X})$ be the set of probability measures on \mathcal{X} . Let X denote a random variable with law P , and $\mathbb{E}_X[f(X)] := \int_{\mathcal{X}} f(x) dP(x)$ be the expectation of a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$. Assume that a p.d. kernel $k(x, y)$, $x, y \in \mathcal{X}$, is measurable. The *kernel mean* of a probability measure $P \in \mathcal{P}(\mathcal{X})$ in the RKHS \mathcal{H} generated by a p.d. kernel $k(\cdot, \cdot)$ is an RKHS element

$$m_P := \mathbb{E}_X[k(\cdot, X)] \in \mathcal{H}. \quad (11)$$

The following holds for the kernel mean w.r.t. the RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. These relations can be found and exploited in many kernel mean papers presented in Section 1.

Proposition 2.1 *Let $m_P, m_Q \in \mathcal{H}$ be the kernel means of probability measures $P, Q \in \mathcal{P}(\mathcal{X})$ in the RKHS \mathcal{H} generated by a p.d. kernel $k(\cdot, \cdot)$, respectively. Then,*

1. $\langle m_P, f \rangle_{\mathcal{H}} = \mathbb{E}_X[f(X)], \quad \forall f \in \mathcal{H}, \quad (\text{expectation property})$
2. $\langle m_P, k(\cdot, x) \rangle_{\mathcal{H}} = m_P(x), \quad x \in \mathbb{R}^d,$
3. $\langle m_P, m_Q \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim P, Y \sim Q}[k(X, Y)].$

Throughout this paper, we assume that a p.d. kernel is bounded, i.e., $\sup_{x \in \mathcal{X}} k(x, x) < \infty$. A bounded kernel guarantees that any kernel mean is bounded, i.e., $m_P(x) < \infty$, $\forall x \in \mathcal{X}$, $P \in \mathcal{P}(\mathcal{X})$ (Sriperumbudur et al., 2010).

Characteristic kernel:

Characteristic kernels constitute an important subclass of bounded measurable p.d. kernels. Let $\mathcal{M}(\mathcal{X}) \subset \mathcal{H}$ denote the image of the kernel mean map $P \rightarrow m_P$ on \mathcal{X} , i.e., $\mathcal{M}(\mathcal{X}) := \{m_P \in \mathcal{H} | P \in \mathcal{P}(\mathcal{X})\}$.

Definition 2.2 (Fukumizu et al., 2004)(Sriperumbudur et al., 2010, Definition 6) *A bounded measurable p.d. kernel k is characteristic if the kernel mean map $\mathcal{P}(\mathcal{X}) \rightarrow \mathcal{M}(\mathcal{X}); P \mapsto m_P$ is injective, i.e., for $P, Q \in \mathcal{P}(\mathcal{X})$, $m_P = m_Q$ implies $P = Q$.*

In this paper, we restrict our attention to the class of nonnegative shift-invariant kernels on the Euclidean space $\mathcal{X} = \mathbb{R}^d$. A function $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ is called *positive-definite* if $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, is a p.d. kernel. Such a kernel $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$ is called a *shift-invariant* p.d. kernel. We say a p.d. kernel k is *nonnegative* if $k(x, y) \geq 0 \forall x, y \in \mathbb{R}^d$. We assume that ψ is integrable, i.e., $\int_{\mathbb{R}^d} \psi(x) dx < \infty$. Without loss of generality, consider ψ to be a probability density function, i.e., $\int_{\mathbb{R}^d} \psi(x) dx = 1$. The normalized form $\psi(0) = 1$ of a positive-definite function is obtained by multiplying by an appropriate constant C . The Gaussian and Laplace kernels are examples of such nonnegative shift-invariant kernels, with positive-definite functions of a Gaussian density and a Laplace density, respectively. The Gaussian and Laplace kernels are known to be characteristic (Fukumizu et al., 2008; Sriperumbudur et al., 2010).

Let $P * Q$, $P, Q \in \mathcal{P}(\mathbb{R}^d)$, denote the convolution of probability measures:

$$(P * Q)(B) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} 1_B(x + y) P(dx) Q(dy), B \in \mathcal{B}(\mathbb{R}^d). \quad (12)$$

If P has a density function d , the convolution $d * Q$ is given by

$$(d * Q)(y) = \int_{\mathbb{R}^d} d(y - x)Q(dx), \quad y \in \mathbb{R}^d. \quad (13)$$

The following lemma holds for the kernel mean when a positive-definite function is given by a probability density function ψ on \mathbb{R}^d .

Lemma 2.3 *Let $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, be a shift-invariant p.d. kernel, where ψ is a probability density function on \mathbb{R}^d . Then, every kernel mean m_P of $P \in \mathcal{P}(\mathbb{R}^d)$ in \mathcal{H} is a probability density function $m_P = \psi * P$ on \mathbb{R}^d .*

Proof The kernel mean m_P has a convolution expression:

$$m_P(y) = \int_{\mathbb{R}^d} k(x, y)dP(x) = \int_{\mathbb{R}^d} \psi(y - x)dP(x) = (\psi * P)(y), \quad y \in \mathbb{R}^d. \quad (14)$$

Since ψ is a probability density function, the kernel mean m_P is absolutely continuous, and is a probability density function on \mathbb{R}^d (Sato, 1999, Lemma 27.1, p. 174):

$$\int_{\mathbb{R}^d} m_P(y)dy = \int \int_{\mathbb{R}^d \times \mathbb{R}^d} \psi(y - x)dydP(x) = 1.$$

■

From Lemma 2.3, when the p.d. kernel is shift-invariant with a probability density function ψ , the kernel mean map $P \mapsto m_P$ maps a probability measure to another probability density function. The set $\mathcal{M}(\mathbb{R}^d)$ of kernel means is a set of probability density functions on \mathbb{R}^d ⁵. When the kernel has the normalized form $\psi(0) = 1$, its kernel mean is obtained by multiplying by an appropriate constant, i.e., Cm_P .

In the class of shift-invariant p.d. kernels on \mathbb{R}^d , a necessary and sufficient condition for a p.d. kernel to be characteristic is given using Bochner's Theorem (Sriperumbudur et al., 2010):

5. Note that since $m_P, m_Q \in \mathcal{M}(\mathbb{R}^d)$ are probability density functions, in addition to the Hilbertian norm $\|\cdot\|_{\mathcal{H}}$, the Kullback-Leibler (KL) divergence can be defined in the set $\mathcal{M}(\mathbb{R}^d)$ of kernel means:

$$KL(m_P || m_Q) = \int m_P(x) \log \frac{m_P(x)}{m_Q(x)} dx, \quad m_P, m_Q \in \mathcal{M}(\mathbb{R}^d). \quad (15)$$

This implies that the set $\mathcal{M}(\mathbb{R}^d)$ can be an object for the information geometry (Amari and Nagaoka, 2000). The kernel mean map $P \mapsto m_P$ maps any m-flat manifold to an m-flat manifold in $\mathcal{M}(\mathbb{R}^d)$, but this is not true for an e-flat manifold.

Potentially, methods relating the KL and other divergences in this context may be used for the kernel mean. Note, however, that the kernel mean estimators in applications (Song et al., 2009; Fukumizu et al., 2013) are given by a form of the weighted sum $\hat{m}_P = \sum_{i=1}^n w_i k(\cdot, X_i)$ with data $\{X_i\}_{i=1}^n$ and associated weights $w \in \mathbb{R}^n$. Weights w are allowed to take negative values, and the kernel mean estimator \hat{m}_P may lie outside $\mathcal{M}(\mathbb{R}^d)$, namely, $\hat{m}_P \in \mathcal{H} \setminus \mathcal{M}(\mathbb{R}^d)$. Then, the KL divergence between the estimators cannot be defined.

Theorem 2.4 (Bochner) *A continuous function $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ is positive-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on \mathbb{R}^d :*

$$\psi(x) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\varpi^\top x} d\Lambda(\varpi), \quad x \in \mathbb{R}^d. \quad (16)$$

Theorem 2.5 (Sriperumbudur et al., 2010, Theorem 9) *Let $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, be a shift-invariant p.d. kernel with a bounded continuous positive-definite function ψ . Then, k is characteristic if and only if the finite nonnegative Borel measure Λ of the Fourier transform of ψ , i.e., (16), has the entire support $\text{supp}(\Lambda) = \mathbb{R}^d$.*

In the next section, Theorem 2.5 is applied to the Lévy–Khintchine formula. This allows us to show that the nonnegative shift-invariant kernel on \mathbb{R}^d is characteristic if ψ is a bounded continuous density of a symmetric infinitely divisible distribution. When a p.d. kernel $k(\cdot, \cdot)$ is on $\mathcal{X} = \mathbb{R}^d$, statement 3 in Proposition 2.1 has the following expression:

Proposition 2.6 *Let $m_P, m_Q \in \mathcal{H}$ be kernel means of $P, Q \in \mathcal{P}(\mathbb{R}^d)$ in \mathcal{H} generated by a p.d. kernel $k(x, y)$, $x, y \in \mathbb{R}^d$. Then,*

$$\langle m_P, m_Q \rangle_{\mathcal{H}} = (\tilde{m}_P * Q)(0) = (\tilde{m}_Q * P)(0) \quad (17)$$

where $\tilde{m}_P(x) := m_P(-x)$, and so is \tilde{m}_Q .

Proof From statement 3 in Proposition 2.1,

$$\mathbb{E}_{X \sim P, Y \sim Q}[k(X, Y)] = \int_{\mathbb{R}^d} m_P(y) dQ(y) = \int_{\mathbb{R}^d} \tilde{m}_P(-y) dQ(y) = (\tilde{m}_P * Q)(0). \quad (18)$$

■

3. Characteristic Kernels and Infinitely Divisible Distributions

This section shows the main result of this paper. Section 3.1 proposes CID kernels in the class of nonnegative shift-invariant kernels on \mathbb{R}^d . The CID kernels are guaranteed to be characteristic. Section 3.2 shows that the kernel mean of an infinitely divisible distribution in the RKHS generated by a CID kernel is infinitely divisible. The kernel mean is described by the generating triplet of the infinitely divisible distribution. The kernel means of the Gaussian distribution, stable distribution, and subclasses of the GH distribution given in Section 4 are special cases of this result. Section 3.3 shows how to generate a new CID kernel from an (asymmetric) infinitely divisible distribution using the symmetrization operation.

3.1 The Convolution Infinitely Divisible Kernels

We begin with the definition of the infinite divisibility of a probability distribution. Let $*_{i=1}^n P_i$ denote the convolution $P_1 * \cdots * P_n$ of probability measures $P_1, \dots, P_n \in \mathcal{P}(\mathbb{R}^d)$. Let P^{*n} denote the n -fold convolution $*_{i=1}^n P$ of a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$.

Definition 3.1 (Sato, 1999, Definition 7.1, p. 31) A probability measure P on \mathbb{R}^d is infinitely divisible if, for any positive integer n , there is a probability measure P_n on \mathbb{R}^d such that $P = P_n^{*n}$.

Every infinitely divisible distribution P on \mathbb{R}^d has a Lévy–Khintchine representation of its characteristic function. Let $\hat{P}(\theta) := \mathbb{E}_X[\exp(i\theta^\top X)]$, $\theta \in \mathbb{R}^d$ denote the characteristic function of a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$. Let $x \wedge y = \min\{x, y\}$, $x, y \in \mathbb{R}$. Let 1_B be the indicator function on \mathbb{R}^d with $B \subset \mathbb{R}^d$.

Theorem 3.2 (Sato, 1999, Theorem 8.1, p. 37) (i) If P is an infinitely divisible distribution on \mathbb{R}^d , the characteristic function $\hat{P}(\theta)$ is given by

$$\hat{P}(\theta) = \exp \left(i\theta^\top \gamma - \frac{1}{2} \theta^\top A \theta + \int_{\mathbb{R}^d} \left(e^{i\theta^\top x} - 1 - i\theta^\top x 1_{\{|x| \leq 1\}}(x) \right) \nu(dx) \right), \quad \theta \in \mathbb{R}^d, \quad (19)$$

where $\gamma \in \mathbb{R}^d$, A is a symmetric nonnegative-definite $d \times d$ matrix, and ν is a measure on \mathbb{R}^d satisfying

$$\nu(\{\mathbf{0}\}) = 0 \quad \text{and} \quad \int_{\mathbb{R}^d} (|x|^2 \wedge 1) \nu(dx) < \infty. \quad (20)$$

(ii) This representation of an infinitely divisible distribution P by (A, ν, γ) is unique.

(iii) Conversely, for any $\gamma \in \mathbb{R}^d$, symmetric nonnegative-definite $d \times d$ matrix A , and measure ν satisfying (20), there exists an infinitely divisible distribution P such that (i) holds.

(A, ν, γ) is called the *generating triplet* of an infinitely divisible distribution P . A is called the covariance matrix of the Gaussian factor of P , and ν is the *Lévy measure* of P .

The support of every infinitely divisible measure is unbounded, except for the point measures $\delta_a(\cdot)$, $a \in \mathbb{R}^d$ (for $B \in \mathcal{B}(\mathbb{R}^d)$, if $a \in B$, $\delta_a(B) = 1$; otherwise, $\delta_a(B) = 0$) (Sato, 1999, Examples 7.2, p. 31). A continuous bounded density function of a symmetric infinitely divisible measure can be used as a p.d. kernel, and this kernel is characteristic, as shown below. Unbounded probability density functions and discrete probability measures are excluded.

A probability measure \tilde{P} is called the *dual* of a probability measure P if $\tilde{P}(B) = P(-B)$ for every $B \in \mathcal{B}(\mathbb{R}^d)$ with $-B := \{-x : x \in B\}$ (Sato, 1999, p.8). A probability measure P is *symmetric* if $P = \tilde{P}$.

Theorem 3.3 Let P be a symmetric and infinitely divisible measure on \mathbb{R}^d . Assume that P has a continuous bounded density function $d_{ID}(x)$, $x \in \mathbb{R}^d$. Then, the function $k_{ID}(x, y) = d_{ID}(x - y)$, $x, y \in \mathbb{R}^d$, is a p.d. kernel and is characteristic.

Proof A probability measure P on \mathbb{R}^d is symmetric if and only if the characteristic function $\hat{P}(\theta)$, $\theta \in \mathbb{R}^d$ is real valued (Sato, 1999, p.67). If P is symmetric and infinitely divisible, $\hat{P}(\theta) > 0$ for every $\theta \in \mathbb{R}^d$ from the Lévy–Khintchine formula (19). Since the Fourier transform \hat{P} of P is a finite nonnegative measure on \mathbb{R}^d , the function $k_{ID}(x, y)$, $x, y \in \mathbb{R}^d$, is a p.d. kernel (Bochner’s theorem). Since the Fourier transform has the entire support \mathbb{R}^d ,

the p.d. kernel $k_{ID}(x, y)$ is characteristic (Theorem 2.5). ■

In this paper, we call $k_{ID}(x, y)$ the *convolution infinitely divisible (CID) kernel*⁶. If the infinite divisibility is clear in the sense of convolution on the positive-definite function, we simply call it an infinitely divisible kernel. The CID kernels include the following examples of infinitely divisible distributions.

Example 3.4 *CID kernels include positive-definite functions of: $S\alpha S$ distributions ($0 < \alpha \leq 2$, $\alpha = 2$ corresponds to Gaussian distributions, $\alpha = 1$ corresponds to Cauchy distributions), symmetric Student's t -distribution (Grosswald, 1976), SGH distributions, and $TS\alpha S$ distributions (Rachev et al., 2011; Rosiński, 2007; Bianchi et al., 2010).*

In Section 4, we examine the details of the CID kernel of the α -stable distribution and the GH distribution. Theorem 3.3 assumes that P has a continuous bounded density function. In general, there is no known necessary and sufficient condition for an infinitely divisible distribution to have a density function (Sato, 1999, p. 177). Some sufficient conditions are known. If the covariance matrix A of the Gaussian factor is full-rank, the infinitely divisible distribution is absolutely continuous, and has a density function. If $A = 0$, i.e., a *purely non-Gaussian infinitely divisible* distribution, sufficient conditions are given in Sato (1999, Theorem 27.7, Theorem 27.10, p. 177). A known fact is that every nondegenerate⁷ self-decomposable distribution on \mathbb{R}^d is absolutely continuous (Sato, 1999, Theorem 27.13, p. 181). See Appendix A.1 for the definition of a self-decomposable distribution. Thus, we obtain the following Corollary:

Corollary 3.5 *Let P be a symmetric and nondegenerate self-decomposable probability measure on \mathbb{R}^d . Assume that the density $d_{SD}(x)$, $x \in \mathbb{R}^d$, is continuous and bounded. The function $k_{SD}(x, y) = d_{SD}(x - y)$, $x, y \in \mathbb{R}^d$, is a p.d. kernel and is characteristic.*

Proof Corollary 3.5 is a special case of Theorem 3.3. ■

The α -stable distribution, Student's t -distribution, and some of the GH distributions are known to be self-decomposable, as well as infinitely divisible (Sato, 1999, p. 98), (Shanbhag and Sreehari, 1979).

3.2 Kernel Mean of Infinitely Divisible Distribution

In this section, we consider the kernel mean of an infinitely divisible distribution. We see that the kernel mean of an infinitely divisible distribution in RKHS with a CID kernel

6. The term *infinite divisibility* on a p.d. kernel is understood in the sense of the usual product: a p.d. kernel $\varphi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ on a nonempty set \mathcal{X} is called *infinitely divisible* if, for each $n \in \mathbb{N}$, there exists a p.d. kernel $\varphi_n : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ such that $\varphi = (\varphi_n)^n$ (Berg et al., 1984, Definition 2.6, p. 76).

To distinguish from this context, we use the term *convolution infinitely divisible*.

7. A measure P on \mathbb{R}^d is said to be *degenerate* if there exist a vector $a \in \mathbb{R}^d$ and a proper linear subspace V of \mathbb{R}^d with $\dim V \leq d - 1$ such that $\text{supp}(P) \subset a + V$ (Sato, 1999, Definition 24.16, p. 156). Otherwise, P is called *nondegenerate*.

is infinitely divisible. This section provides a general perspective of the kernel mean, and Section 4 gives specific examples of this. For instance, see Section 4.1 for one of the simplest cases, namely Gaussians. This section is a generalization of the Gaussian case.

Let $P_{ID}(\cdot; A, \nu, \gamma)$ be an infinitely divisible distribution on \mathbb{R}^d with a generating triplet (A, ν, γ) , and, if it exists, density $d_{ID}(x; A, \nu, \gamma)$. Let $I(\mathbb{R}^d)$ be the set of infinitely divisible distributions on \mathbb{R}^d . An infinitely divisible distribution $P_{ID}(\cdot; A, \nu, \gamma)$ is symmetric if and only if $\gamma = 0$ and the Lévy measure ν is symmetric (Sato, 1999, p.114).

Let $k_{(A, \nu_s)}(x, y)$, $x, y \in \mathbb{R}^d$, be the CID kernel with a generating triplet $(A, \nu_s, 0)$, where ν_s is a symmetric Lévy measure. There is the unique RKHS \mathcal{H}_{A, ν_s} associated with the CID kernel $k_{(A, \nu_s)}(x, y)$. The following holds for the infinitely divisible kernel mean:

Proposition 3.6 *Let \mathcal{H}_{A_0, ν_s} be the RKHS with a CID kernel $k_{A_0, \nu_s}(x, y)$, where A_0 is a $d \times d$ nonnegative definite matrix and ν_s is a symmetric Lévy measure. Then, the kernel mean m_P of $P_{ID}(\cdot; A, \nu, \gamma) \in I(\mathbb{R}^d)$ in \mathcal{H}_{A_0, ν_s} is*

$$m_P = d_{ID}(x; A_0 + A, \nu_s + \nu, \gamma). \quad (21)$$

Proof By definition, for $y \in \mathbb{R}^d$,

$$\begin{aligned} m_P(y) &= \int_{\mathbb{R}^d} k_{A, \nu_s}(x, y) dP(x) = (d_{ID}(x; A_0, \nu_s, 0) * P_{ID}(\cdot; A, \nu, \gamma))(y) \\ &= d_{ID}(x; A_0 + A, \nu_s + \nu, \gamma). \end{aligned} \quad (22)$$

For the third equality, we used that the set $I(\mathbb{R}^d)$ is closed under convolution:

Lemma 3.7 *Let $P_{ID}(\cdot; A_i, \nu_i, \gamma_i)$, $i = 1, \dots, n$, be infinitely divisible distributions on \mathbb{R}^d . The convolution is also an infinitely divisible distribution:*

$$\bigstar_{i=1}^n P_{ID}(\cdot; A_i, \nu_i, \gamma_i) = P_{ID}(\cdot; \sum_{i=1}^n A_i, \sum_{i=1}^n \nu_i, \sum_{i=1}^n \gamma_i). \quad (23)$$

Considering the special case of $n = 2$, $(A_1, \nu_1, \gamma_1) = (A_0, \nu_s, 0)$, and $(A_2, \nu_2, \gamma_2) = (A, \nu, \gamma)$ in Lemma 3.7, we reach (21). Since $d_{ID}(x; A_0, \nu_s, 0)$ is a density function, $P_{ID}(x; A_0 + A, \nu_s + \nu, \gamma)$ is absolutely continuous, and has the density $d_{ID}(x; A_0 + A, \nu_s + \nu, \gamma)$ (Sato, 1999, Lemma 27.1, p. 174). \blacksquare

The image of the CID kernel mean map of $I(\mathbb{R}^d)$ is a set of infinitely divisible densities. The following holds for the kernel mean w.r.t. the RKHS inner product. The inner product becomes an infinitely divisible density value.

Proposition 3.8 *Let $m_P, m_Q \in \mathcal{H}_{A_0, \nu_s}$ be kernel means of infinitely divisible distributions $P_{ID}(\cdot; A_P, \nu_P, \gamma_P)$ and $Q_{ID}(\cdot; A_Q, \nu_Q, \gamma_Q)$, respectively, in the same RKHS \mathcal{H}_{A_0, ν_s} generated by a CID kernel k_{A_0, ν_s} . Then,*

$$1. \quad \langle m_P, k_{A_0, \nu_s}(\cdot, x) \rangle_{\mathcal{H}_{A_0, \nu_s}} = d_{ID}(x; A_0 + A, \nu_s + \nu, \gamma_P), \quad x \in \mathbb{R}^d,$$

$$2. \langle m_P, m_Q \rangle_{\mathcal{H}_{A_0, \nu_s}} = d_{ID}(0; A_0 + A_P + A_Q, \nu_s + \tilde{\nu}_P + \nu_Q, \gamma_Q - \gamma_P) \text{ }^8,$$

where $\tilde{\nu}_P$ is the dual of the Lévy measure ν_P , i.e., $\tilde{\nu}_P(B) = \nu_P(-B)$, $\forall B \in \mathcal{B}(\mathbb{R}^d)$.

Proof Statement 1 is straightforward from Proposition 2.1 and (21). Statement 2 is obtained using Proposition 2.6, the dual of the kernel mean $\tilde{m}_P = d_{ID}(\cdot; A_0 + A, \nu_s + \tilde{\nu}_P, -\gamma_P)$, and Lemma 3.7. \blacksquare

Proposition 3.8 is the most general result for the inner product (1) of the kernel mean. Specific examples of the inner product in the cases of the Gaussian distribution, stable distribution, and subclasses of the GH distribution are given in Section 4. We now mention the smoothness of the CID kernel and the kernel mean of an infinitely divisible distribution. The rapid decrease in $\hat{P}(\theta)$ as $|\theta| \rightarrow \infty$ implies the smoothness of the density $d(x)$ (Sato, 1999, p. 189). Sufficient conditions on $\hat{P}(\theta)$ and the Lévy measure ν such that the density becomes a C^n function ($n = 0, 1, \dots, \infty$) are known in Sato (1999, Proposition 28.1, 28.3, p. 190).

3.3 Generation of Convolution Infinitely Divisible Kernels

This section describes how to generate a new CID kernel from a continuous bounded infinitely divisible density $d_{ID}(x; A, \nu, \gamma)$ (not necessarily a symmetric density) by the symmetrization operation.

The *symmetrization* P^* of a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$ is $P^* = P * \tilde{P}$, where \tilde{P} is the dual of P . P^* is symmetric. If P is infinitely divisible, so is P^* ⁹. By the symmetrization $P \mapsto P^*$, the generating triplet $(A, \nu, \gamma) \mapsto (2A, 2\nu_s, 0)$, where ν_s is a symmetric Lévy measure with $\nu_s(B) = \frac{\nu(B) + \nu(-B)}{2}$ for $B \in \mathcal{B}(\mathbb{R}^d)$ (Sato, 1999, p. 114).

The following example shows that the Laplace kernel is a CID kernel generated from an exponential distribution by symmetrization.

Example 3.9 (Sato, 1999, Example 15.14, p. 98) Let P be an exponential distribution with the density $d(x) = \lambda \exp(-\lambda x) 1_{[0, \infty)}(x)$ on \mathbb{R} . The exponential distribution is infinitely divisible. The dual \tilde{P} has density $\tilde{d}(x) = \lambda \exp(\lambda x) 1_{(-\infty, 0)}(x)$. The characteristic function of the symmetrization P^* is

$$\hat{P}^*(\theta) = \hat{P}(\theta) \hat{\tilde{P}}(\theta) = \frac{\lambda}{\lambda - i\theta} \cdot \frac{\lambda}{\lambda + i\theta} = \frac{\lambda^2}{\lambda^2 + \theta^2}. \quad (24)$$

This is the characteristic function of the Laplace distribution with density $d^*(x) = \frac{\lambda}{2} \exp(-\lambda|x|)$ on \mathbb{R} . The Laplace kernel $k^*(x, y) = d^*(x - y)$ is a CID kernel generated from an exponential distribution P by the symmetrization P^* , and $k^*(x, y)$ is characteristic.

8. Since P and Q are exchangeable, it also holds that $\langle m_P, m_Q \rangle_{\mathcal{H}_{A_0, \nu_s}} = d_{ID}(0; A_0 + A_P + A_Q, \nu_s + \nu_P + \tilde{\nu}_Q, \gamma_P - \gamma_Q)$.

9. On the other hand, let P_s be a symmetric measure of a probability measure P defined by $P_s(B) = \frac{P(B) + P(-B)}{2}$ for $B \in \mathcal{B}(\mathbb{R}^d)$. If P is infinitely divisible, P_s is not generally infinitely divisible (Sato, 1999, p. 67).

4. Subclasses of Convolution Infinitely Divisible Kernels

This section explores subclasses of the CID kernels. Section 4.1 gives one of the simplest cases, Gaussians. It is well known that the Gaussian kernel is characteristic (Fukumizu et al., 2008; Sriperumbudur et al., 2010), and the closed-form solution of the kernel mean of the Gaussian distribution is used in, e.g., Smola et al. (2007), Song et al. (2008), Muandet et al. (2012), and McCalman et al. (2013). Section 4.1 will help our understanding of later sections on the stable distributions and GH distributions. These are generalizations of the Gaussian case, and similar arguments can be applied. Section 4.2 explores the univariate α -stable distribution on \mathbb{R} . Section 4.3 explores the multivariate α -stable distribution on \mathbb{R}^d , including the sub-Gaussian α -stable distribution and the isotropic α -stable distribution. Section 4.4 explores the multivariate GH distribution on \mathbb{R}^d .

4.1 Gaussian Kernel on \mathbb{R}^d (Gaussian Distribution)

Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ be a d -dimensional Gaussian random vector with a mean vector $\mu \in \mathbb{R}^d$ and a nondegenerate (rank d) covariance matrix $R \in \mathbb{R}^{d \times d}$. Then, we write $X \sim N(\mu, R)$. We write $P_G(\cdot; \mu, R)$ and $d_G(x; \mu, R)$ for the distribution and density, respectively.

Proposition 4.1 *The characteristic function of a d -dimensional Gaussian random vector $X \sim N(\mu, R)$ is given by*

$$\hat{P}(\theta) = \exp\left(-\frac{1}{2}\theta^\top R\theta + i\theta^\top \mu\right), \quad \theta \in \mathbb{R}^d. \quad (25)$$

The Gaussian distribution is symmetric if and only if the mean vector $\mu = 0$.

Theorem 4.2 *Let R be a $d \times d$ positive-definite matrix. Let $k_R : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be the function defined by a symmetric Gaussian density $k_R(x, y) = d_G(x - y; 0, R)$. Then, the function $k_R(\cdot, \cdot)$ is a bounded p.d. kernel and is characteristic.*

Proof It is a well-known fact that the Gaussian kernel is a characteristic p.d. kernel (Fukumizu et al., 2008; Sriperumbudur et al., 2010). Theorem 4.2 is a special case of Theorem 3.3. From Proposition 4.1, since the Fourier transform of the density $d_G(x; 0, R)$ is a finite nonnegative measure on \mathbb{R}^d , the function $k_R(x, y)$ is a p.d. kernel (Bochner's theorem). Since the Fourier transform has the entire support \mathbb{R}^d , the Gaussian kernel $k_R(x, y)$ is characteristic (Theorem 2.5). \blacksquare

A p.d. kernel $k_R(\cdot, \cdot)$ uniquely determines an RKHS \mathcal{H}_R , and \mathcal{H}_R is a characteristic RKHS. We consider the kernel mean of a multivariate Gaussian distribution. When the RKHS is chosen as a Gaussian RKHS \mathcal{H}_R , the kernel mean has a simple form, a multivariate Gaussian density.

Proposition 4.3 *The kernel mean m_P of a d -dimensional Gaussian distribution $P_G(\cdot; \mu, R)$ in a Gaussian RKHS \mathcal{H}_{R_0} , where R_0 is a nonsingular $d \times d$ covariance matrix, has the form:*

$$m_P = d_G(\cdot; \mu, R_0 + R). \quad (26)$$

Proof By definition, for $y \in \mathbb{R}^d$,

$$m_P(y) = \int_{\mathbb{R}^d} k_R(x, y) dP(x) = (d_G(x; 0, R_0) * P_G(\cdot; \mu, R))(y) = d_G(y; \mu, R_0 + R). \quad (27)$$

For the third equality, we used the following Gaussian convolution property:

Lemma 4.4 *Let $P_G(\cdot; \mu_i, R_i)$, $i = 1, \dots, n$, be n d -dimensional Gaussian distributions. Then, the n -th convolution is the Gaussian*

$$\bigstar_{i=1}^n P_G(\cdot; \mu_i, R_i) = P_G(\cdot; \sum_{i=1}^n \mu_i, \sum_{i=1}^n R_i). \quad (28)$$

Letting $n = 2$, $(\mu_1, R_1) = (0, R_0)$, and $(\mu_2, R_2) = (\mu, R)$ in Lemma 4.4, we obtain (26). \blacksquare

The following holds for the Gaussian kernel mean w.r.t. the RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{R_0}}$. The inner product becomes a Gaussian density value.

Proposition 4.5 *Let $m_P, m_Q \in \mathcal{H}_{R_0}$ be kernel means of multivariate Gaussian distributions $P_G(\cdot; \mu_P, R_P)$ and $Q_G(\cdot; \mu_Q, R_Q)$, respectively, in the Gaussian RKHS \mathcal{H}_{R_0} generated by a Gaussian kernel k_{R_0} . Then,*

1. $\langle m_P, k_{R_0}(\cdot, x) \rangle_{\mathcal{H}_{R_0}} = d_G(x; \mu_P, R_0 + R_P), \quad x \in \mathbb{R}^d,$
2. $\langle m_P, m_Q \rangle_{\mathcal{H}_{R_0}} = d_G(0; \mu_Q - \mu_P, R_0 + R_P + R_Q).$

Proof Statement 1 is straightforward from Proposition 2.1 and (26). Statement 2 is obtained using Proposition 2.6, the dual of the kernel mean $\tilde{m}_P = d_G(\cdot; -\mu_P, R_0 + R_P)$, and Lemma 4.4. \blacksquare

4.2 α -Stable Kernel on \mathbb{R} (Univariate α -Stable Distribution)

We now consider the univariate α -stable distribution ($0 < \alpha \leq 2$) on \mathbb{R} . $\alpha = 2$ corresponds to the Gaussian distribution. While the Gaussian distribution has light tails, the non-Gaussian stable distributions ($0 < \alpha < 2$) have heavy tails (when $0 < \alpha < 1$ and $\beta = \pm 1$, the stable distribution is one-sided, otherwise two-sided). The definition is as follows:

Definition 4.6 (Samorodnitsky and Taqqu, 1994, Definition 1.1.1, p. 2) *An \mathbb{R} -valued random variable X is said to have a stable distribution if, for any positive numbers A and B , there is a positive number C and a real number D such that*

$$AX^{(1)} + BX^{(2)} \stackrel{d}{=} CX + D,$$

where $X^{(1)}$ and $X^{(2)}$ are independent copies of X , and $\stackrel{d}{=}$ denotes equality in the distribution.

See Samorodnitsky and Taqqu (1994, Section 1.1, p.2) for other (necessary and sufficient) definitions. For any stable random variable X , there is a number $\alpha \in (0, 2]$ such that the number C satisfies $C^\alpha = A^\alpha + B^\alpha$ (Samorodnitsky and Taqqu, 1994, Theorem 1.1.2, p. 3). The number α is called the *characteristic exponent*. A stable random variable X with index α is called α -stable. The characteristic function of the α -stable random variable is given as follows:

Theorem 4.7 (Samorodnitsky and Taqqu, 1994, Definition 1.1.6, p. 5) *A random variable X is α -stable ($\alpha \in (0, 2]$) in \mathbb{R} if and only if there are parameters $\sigma \geq 0$, $\beta \in [-1, 1]$, and $\mu \in \mathbb{R}$ such that its characteristic function has the form*

$$\hat{P}(\theta) = \begin{cases} \exp(-\sigma^\alpha |\theta|^\alpha (1 - i\beta(\operatorname{sgn}\theta) \tan \frac{\pi\alpha}{2}) + i\mu\theta) & (\alpha \neq 1), \\ \exp(-\sigma |\theta| (1 + i\beta \frac{2}{\pi}(\operatorname{sgn}\theta) \ln |\theta|) + i\mu\theta) & (\alpha = 1), \end{cases}$$

where $\operatorname{sgn}\theta$ is a sign function

$$\operatorname{sgn}\theta = \begin{cases} 1 & \theta > 0, \\ 0 & \theta = 0, \\ -1 & \theta < 0. \end{cases}$$

When $\alpha \in (0, 2)$, the parameters σ , β , and μ are unique. When $\alpha = 2$, β is irrelevant, and σ and μ are unique.

An α -stable random variable X in \mathbb{R} is specified by a parameter triplet (σ, β, μ) , and we write $X \sim S_\alpha(\sigma, \beta, \mu)$. σ is a scale parameter, β is a skewness parameter, and μ is a location parameter. When $\sigma = 0$, the stable random variable X is a degenerate random variable (a point measure $\delta_a(x)$, $a \in \mathbb{R}$). Let $P_{\text{stable}}(\cdot; \alpha, \sigma, \beta, \mu)$ and $d_{\text{stable}}(x; \alpha, \sigma, \beta, \mu)$ denote the distribution and density, respectively. It is known that every nondegenerate stable distribution has a C^∞ density function (Sato, 1999, Example 28.2, p. 190). In general, it is known that an α -stable density does not have a closed-form solution, except for some special cases. The known closed-form solution of a univariate α -stable density is given in Appendix A.2.

We consider the p.d. kernel generated by an α -stable density ($0 < \alpha \leq 2$). When $\alpha \in (0, 2)$, an α -stable random variable $X \sim S_\alpha(\sigma, \beta, \mu)$ is symmetric α -stable (S α S) if and only if $\beta = \mu = 0$ (Samorodnitsky and Taqqu, 1994, Property 1.2.5, p. 11). A 2-stable random variable is symmetric 2-stable (S2S) if and only if $\mu = 0$.

Theorem 4.8 *Let $\alpha \in (0, 2]$ and $\sigma > 0$. Let $k_{\alpha, \sigma} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be the function defined by a nondegenerate S α S density with a scale parameter σ , as $k_{\alpha, \sigma}(x, y) = d_{\text{stable}}(x - y; \alpha, \sigma, 0, 0)$. The function $k_{\alpha, \sigma}(\cdot, \cdot)$ is a bounded continuous p.d. kernel and is characteristic.*

Proof Theorem 4.8 is a special case of Theorem 3.3. Since $d_{\text{stable}}(x; \alpha, \sigma, 0, 0)$ is the density of a nondegenerate S α S two-sided distribution, $d_{\text{stable}}(x; \alpha, \sigma, 0, 0)$ is a C^∞ function and is bounded. The positive-definiteness and the characteristic property of $k_{\alpha, \sigma}(x, y)$ are shown from the following Lemma:

Lemma 4.9 *Let $\alpha \in (0, 2]$. X is a nondegenerate S α S random variable in \mathbb{R} if and only if there is a parameter $\sigma > 0$ such that its characteristic function has the form:*

$$\hat{P}(\theta) = \exp(-\sigma^\alpha |\theta|^\alpha).$$

From Lemma 4.9, since the Fourier transform of $d_{\text{stable}}(x; \alpha, \sigma, 0, 0)$ is a finite nonnegative measure on \mathbb{R} , the function $k_{\alpha, \sigma}(x, y)$ is a p.d. kernel (Bochner's theorem). Since the Fourier transform has the entire support \mathbb{R} , the p.d. kernel $k_{\alpha, \sigma}(x, y)$ is characteristic (Theorem 2.5). \blacksquare

A p.d. kernel $k_{\alpha, \sigma}(\cdot, \cdot)$ uniquely determines an RKHS $\mathcal{H}_{\alpha, \sigma}$, and $\mathcal{H}_{\alpha, \sigma}$ is a characteristic RKHS. We consider a kernel mean of a univariate α -stable distribution. When the RKHS is chosen as the α -stable RKHS $\mathcal{H}_{\alpha, \sigma}$, the kernel mean has a simple form, an α -stable density.

Proposition 4.10 *Let $P_{\text{stable}}(\cdot; \alpha, \sigma, \beta, \mu)$ be an α -stable distribution on \mathbb{R} with $\alpha \in (0, 2]$, $\sigma \geq 0$, $\beta \in [-1, 1]$, and $\mu \in \mathbb{R}$. The kernel mean m_P of $P_{\text{stable}}(\cdot; \alpha, \sigma, \beta, \mu)$ in $\mathcal{H}_{\alpha, \sigma_0}$ ($\sigma_0 > 0$) is given by*

$$m_P = d_{\text{stable}}(\cdot; \alpha, \|(\sigma_0, \sigma)\|_{\alpha}, \frac{\sigma^{\alpha} \beta}{\|(\sigma_0, \sigma)\|_{\alpha}^{\alpha}}, \mu), \quad (29)$$

where $\|(\sigma_0, \sigma)\|_{\alpha} := (\sigma_0^{\alpha} + \sigma^{\alpha})^{\frac{1}{\alpha}} > 0$.

Proof By definition, for $y \in \mathbb{R}$,

$$\begin{aligned} m_P(y) &= \int_{\mathbb{R}} k_{\alpha, \sigma_0}(x, y) dP(x) = (d_{\text{stable}}(x; \alpha, \sigma_0, 0, 0) * P_{\text{stable}}(\cdot; \alpha, \sigma, \beta, \mu))(y) \\ &= d_{\text{stable}}(y; \alpha, \|(\sigma_0, \sigma)\|_{\alpha}, \frac{\sigma^{\alpha} \beta}{\|(\sigma_0, \sigma)\|_{\alpha}^{\alpha}}, \mu). \end{aligned} \quad (30)$$

For the third equality, we used the following convolution stability property:

Property 4.11 (Samorodnitsky and Taqqu, 1994, Property 1.2.1, p. 10) *Let $P_{\text{stable}}(\cdot; \alpha, \sigma_i, \beta_i, \mu_i)$, $i = 1, \dots, d$, be d univariate α -stable distributions. Then, the d -th convolution is given by*

$$\bigstar_{i=1}^d P_{\text{stable}}(\cdot; \alpha, \sigma_i, \beta_i, \mu_i) = P_{\text{stable}}(\cdot; \alpha, \|\sigma\|_{\alpha}, \frac{1}{\|\sigma\|_{\alpha}^{\alpha}} \sum_{i=1}^d \beta_i \sigma_i^{\alpha}, \sum_{i=1}^d \mu_i), \quad (31)$$

where $\|\sigma\|_{\alpha} = (\sum_{i=1}^d \sigma_i^{\alpha})^{\frac{1}{\alpha}}$ for a scale vector $\sigma = (\sigma_1, \dots, \sigma_d)$, the convoluted skewness is given by the weighted average of $\{\beta_i\}$ with respect to $\{\sigma_i^{\alpha}\}$, and μ is the sum of locations $\{\mu_i\}$.

Letting $d = 2$, $(\sigma_1, \beta_1, \mu_1) = (\sigma_0, 0, 0)$, and $(\sigma_2, \beta_2, \mu_2) = (\sigma, \beta, \mu)$ in Property 4.11, (29) is obtained. Even when $P_{\text{stable}}(\cdot; \alpha, 0, \beta, \mu) = \delta_{\mu}(\cdot)$, (29) holds. \blacksquare

The following holds for the α -stable kernel mean w.r.t. the α -stable RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\alpha, \sigma_0}}$.

Proposition 4.12 *Let $\alpha \in (0, 2]$ and $\sigma_0 > 0$. Let $m_P, m_Q \in \mathcal{H}_{\alpha, \sigma_0}$ be kernel means of stable distributions $P_{\text{stable}}(\cdot; \alpha, \sigma_P, \beta_P, \mu_P)$ and $Q_{\text{stable}}(\cdot; \alpha, \sigma_Q, \beta_Q, \mu_Q)$ in the α -stable RKHS $\mathcal{H}_{\alpha, \sigma_0}$ generated by an α -stable kernel $k_{\alpha, \sigma_0}(\cdot, \cdot)$, respectively. Then,*

$$1. \langle m_P, k_{\alpha, \sigma_0}(\cdot, x) \rangle_{\mathcal{H}_{\alpha, \sigma_0}} = d_{\text{stable}}(x; \alpha, \|(\sigma_0, \sigma_P)\|_{\alpha}, \frac{\sigma_P^{\alpha} \beta_P}{\|(\sigma_0, \sigma_P)\|_{\alpha}^{\alpha}}, \mu_P), \quad x \in \mathbb{R}^d,$$

$$2. \langle m_P, m_Q \rangle_{\mathcal{H}_{\alpha, \sigma_0}} = d_{\text{stable}}(0; \alpha, \|(\sigma_0, \sigma_P, \sigma_Q)\|_{\alpha}, \frac{\sigma_Q^{\alpha} \beta_Q - \sigma_P^{\alpha} \beta_P}{\|(\sigma_0, \sigma_P, \sigma_Q)\|_{\alpha}^{\alpha}}, \mu_Q - \mu_P).$$

Proof Statement 1 is straightforward from Proposition 2.1 and (29). Statement 2 is obtained using Proposition 2.6, the dual of the kernel mean $\tilde{m}_P = d_{\text{stable}}(\cdot; \alpha, \|(\sigma_0, \sigma_P)\|_{\alpha}, -\frac{\sigma_P^{\alpha} \beta_P}{\|(\sigma_0, \sigma_P)\|_{\alpha}^{\alpha}}, -\mu_P)$, and Property 4.11. ■

From the results of this section, the α -stable kernel, the kernel mean of the α -stable distribution, and the relevant inner product involve the α -stable density. In general, the α -stable density is known not to have a closed-form solution, except for some special cases. Appendix A.2 shows instances of known closed-form solutions for a univariate α -stable density, expressed by elementary functions and special functions. Note that the α -stable kernel, α -stable kernel mean, and inner product above may be given by such closed-form expressions. The characteristic exponent $\alpha \in (0, 2]$ must be identical for the embedded α -stable distribution and the α -stable RKHS. Embedding the α -stable distribution in the conjugate RKHS (the same characteristic exponent $\alpha \in (0, 2]$), the kernel mean and the relevant inner product are given by the α -stable density form.

4.3 α -Stable Kernel on \mathbb{R}^d (Multivariate α -Stable Distribution)

In this section, we extend the previous results on univariate α -stable distributions to multivariate α -stable distributions on \mathbb{R}^d . The definition is a straightforward extension of the univariate case:

Definition 4.13 (*Samorodnitsky and Taqqu, 1994, Definition 2.1.1, p. 57*) *A random vector $X = (X_1, \dots, X_d)$ is said to be a stable random vector in \mathbb{R}^d if, for any positive numbers A and B , there is a positive number C and a vector $D \in \mathbb{R}^d$ such that*

$$AX^{(1)} + BX^{(2)} \stackrel{d}{=} CX + D,$$

where $X^{(1)}$ and $X^{(2)}$ are independent copies of X .

See Samorodnitsky and Taqqu (1994, Section 2.1, p.57) for other (necessary and sufficient) definitions. For any stable random vector X , there is a constant $\alpha \in (0, 2]$ such that $C = (A^{\alpha} + B^{\alpha})^{\frac{1}{\alpha}}$ (Samorodnitsky and Taqqu, 1994, Theorem 2.1.2, p. 58). α is called the *characteristic exponent*. The stable random vector X with a characteristic exponent $\alpha \in (0, 2]$ is called α -stable. The characteristic function $\hat{P}(\theta)$ of the α -stable random vector in \mathbb{R}^d is given as follows:

Theorem 4.14 (*Samorodnitsky and Taqqu, 1994, Theorem 2.3.1, p. 65*) *Let $\alpha \in (0, 2)$. Then, $X = (X_1, \dots, X_d)$ is an α -stable random vector in \mathbb{R}^d if and only if there exists a finite measure Γ on the unit sphere $S_{d-1} = \{s \in \mathbb{R}^d : \|s\| = 1\}$ and a vector $\mu^0 \in \mathbb{R}^d$ such that*

$$\hat{P}(\theta) = \begin{cases} \exp\left(-\int_{S_{d-1}} |\theta^{\top} s|^{\alpha} \left(1 - i \operatorname{sgn}(\theta^{\top} s) \tan \frac{\pi\alpha}{2}\right) \Gamma(ds) + i\theta^{\top} \mu^0\right), & (\alpha \neq 1). \\ \exp\left(-\int_{S_{d-1}} |\theta^{\top} s|^{\alpha} \left(1 + i \frac{2}{\pi} \operatorname{sgn}(\theta^{\top} s) \ln |\theta^{\top} s|\right) \Gamma(ds) + i\theta^{\top} \mu^0\right), & (\alpha = 1). \end{cases}$$

The pair (Γ, μ^0) is unique.

The measure Γ is called the *spectral measure*. Examples of the spectral measure Γ are given in Samorodnitsky and Taqqu (1994, Section 2.3), including the discrete measure and the uniform measure. It is known that any nondegenerate stable distribution on \mathbb{R}^d has a C^∞ density function (Sato, 1999, Example 28.2, p. 190). Let $P_{\text{stable}}(x; \alpha, \Gamma, \mu^0)$ denote the α -stable distribution with (Γ, μ^0) and density $d_{\text{stable}}(x; \alpha, \Gamma, \mu^0)$. Let $\mathbb{S}_\alpha(\mathbb{R}^d)$ denote the set of α -stable distributions on \mathbb{R}^d . An α -stable random vector X is called symmetric α -stable (SaS) if $X \stackrel{d}{=} -X$. An SaS random vector is if and only if $\mu^0 = 0$ and Γ is a symmetric measure on S_{d-1} (i.e., $\Gamma(A) = \Gamma(-A)$ for any $A \in \mathcal{B}(S_{d-1})$) (Samorodnitsky and Taqqu, 1994, p.73). The following shows that the multivariate α -stable kernel on \mathbb{R}^d is characteristic:

Theorem 4.15 *Let $\alpha \in (0, 2)$. Let $d_{\text{stable}}(x; \alpha, \Gamma_s, 0)$ be the density of a nondegenerate SaS distribution on \mathbb{R}^d with a symmetric spectral measure Γ_s . The function $k_{\alpha, \Gamma_s}(x, y) = d_{\text{stable}}(x - y; \alpha, \Gamma_s, 0)$, $x, y \in \mathbb{R}^d$, is a bounded continuous p.d. kernel and is characteristic.*

Proof Theorem 4.15 is a special case of Theorem 3.3. Since $d_{\text{stable}}(x; \alpha, \Gamma_s, 0)$ is the density of a nondegenerate SaS distribution on \mathbb{R}^d , $d_{\text{stable}}(x; \alpha, \Gamma_s, 0)$ is a C^∞ function and is bounded. The following Lemma 4.16 shows that $k_{\alpha, \Gamma_s}(x, y) = d_{\text{stable}}(x - y; \alpha, \Gamma_s, 0)$ is p.d. and characteristic.

Lemma 4.16 (Samorodnitsky and Taqqu, 1994, Theorem 2.4.3, p. 73) *X is an SaS vector in \mathbb{R}^d with $0 < \alpha < 2$ if and only if there exists a unique symmetric finite measure Γ_s on the unit sphere S_{d-1} such that*

$$\mathbb{E} \exp \left[i \theta^\top X \right] = \exp \left(- \int_{S_{d-1}} |\theta^\top s|^\alpha \Gamma_s(ds) \right).$$

Γ_s is the spectral measure of the SaS random vector X .

Since the Fourier transform of $d_{\text{stable}}(x; \alpha, \Gamma_s, 0)$ is a finite nonnegative measure on \mathbb{R}^d , $k_{\alpha, \Gamma_s}(x, y)$ is a p.d. kernel (Bochner's theorem). Since the Fourier transform has the entire support \mathbb{R}^d , the p.d. kernel $d_{\text{stable}}(x; \alpha, \Gamma_s, 0)$ is characteristic (Theorem 2.5). \blacksquare

A p.d. kernel $k_{\alpha, \Gamma_s}(\cdot, \cdot)$ uniquely determines an RKHS $\mathcal{H}_{\alpha, \Gamma_s}$, and $\mathcal{H}_{\alpha, \Gamma_s}$ is a characteristic RKHS. We consider the kernel mean of the multivariate α -stable distribution. When the RKHS is chosen as the α -stable RKHS $\mathcal{H}_{\alpha, \Gamma_s}$, the kernel mean results in a multivariate α -stable density.

Proposition 4.17 *Let $\alpha \in (0, 2)$. Let $P_{\text{stable}}(\cdot; \alpha, \Gamma, \mu)$ be a multivariate α -stable distribution on \mathbb{R}^d with a spectral measure Γ and shift vector μ . The kernel mean m_P of $P_{\text{stable}}(\cdot; \alpha, \Gamma, \mu)$ in $\mathcal{H}_{\alpha, \Gamma_s}$ is given by*

$$m_P = d_{\text{stable}}(\cdot; \alpha, \Gamma_s + \Gamma, \mu). \quad (32)$$

Proof By definition, for $y \in \mathbb{R}^d$,

$$\begin{aligned} m_P(y) &= \int_{\mathbb{R}^d} k_{\alpha, \Gamma_s}(x, y) dP(x) = (d_{\text{stable}}(\cdot; \alpha, \Gamma_s, 0) * P_{\text{stable}}(\cdot; \alpha, \Gamma, \mu))(y) \\ &= d_{\text{stable}}(y; \alpha, \Gamma_s + \Gamma, \mu). \end{aligned} \quad (33)$$

For the third equality, we have used that the set $\mathbb{S}_\alpha(\mathbb{R}^d)$ is closed under convolution:

Lemma 4.18 *Let $P_{\text{stable}}(\cdot; \alpha, \Gamma_i, \mu_i)$, $i = 1, 2$, be two multivariate α -stable distributions on \mathbb{R}^d . Then,*

$$*_{i=1}^2 P_{\text{stable}}(\cdot; \alpha, \Gamma_i, \mu_i) = P_{\text{stable}}(\cdot; \alpha, \Gamma_1 + \Gamma_2, \mu_1 + \mu_2). \quad (34)$$

Putting $(\Gamma_1, \mu_1) = (\Gamma_s, 0)$ and $(\Gamma_2, \mu_2) = (\Gamma, \mu)$ in Lemma 4.18, Proposition 4.17 is obtained. Since $P_{\text{stable}}(\cdot; \alpha, \Gamma_s, 0)$ is absolutely continuous, $P_{\text{stable}}(\cdot; \alpha, \Gamma_s + \Gamma, \mu)$ is absolutely continuous, and has density $d_{\text{stable}}(y; \alpha, \Gamma_s + \Gamma, \mu)$ (Sato, 1999, Lemma 27.1, p. 174). ■

The following holds for the multivariate α -stable kernel mean w.r.t. the RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\alpha, \Gamma_s}}$.

Proposition 4.19 *Let $m_P, m_Q \in \mathcal{H}_{\alpha, \Gamma_s}$ be kernel means of multivariate α -stable distributions $P_{\text{stable}}(\cdot; \alpha, \Gamma_P, \mu_P), Q_{\text{stable}}(\cdot; \alpha, \Gamma_Q, \mu_Q) \in \mathbb{S}_\alpha(\mathbb{R}^d)$, respectively, in the RKHS $\mathcal{H}_{\alpha, \Gamma_s}$ generated by a multivariate α -stable kernel k_{α, Γ_s} . Then,*

1. $\langle m_P, k_{\alpha, \Gamma_s}(\cdot, x) \rangle_{\mathcal{H}_{\alpha, \Gamma_s}} = d_{\text{stable}}(x; \alpha, \Gamma_s + \Gamma_P, \mu_P), \quad x \in \mathbb{R}^d,$
2. $\langle m_P, m_Q \rangle_{\mathcal{H}_{\alpha, \Gamma_s}} = d_{\text{stable}}(0; \alpha, \Gamma_s + \tilde{\Gamma}_P + \Gamma_Q, \mu_Q - \mu_P),$

where $\tilde{\Gamma}_P$ is the dual of the spectral measure Γ_P , i.e., $\tilde{\Gamma}_P(B) = \Gamma_P(-B)$, $\forall B \in \mathcal{B}(S^{d-1})$.

Proof Statement 1 is straightforward from Proposition 2.1 and (32). Statement 2 is obtained using Proposition 2.6, the dual of the kernel mean $\tilde{m}_P = d_{\text{stable}}(\cdot; \alpha, \Gamma_s + \tilde{\Gamma}_P, -\mu_P)$, and (34). ■

The kernel mean of a multivariate α -stable distribution in $\mathcal{H}_{\alpha, \Gamma_s}$ is also a multivariate α -stable density. Each computation of the multivariate α -stable kernel, kernel mean (32), and inner product (Proposition 4.19) is a computation of a multivariate α -stable density with a different spectral measure. The general class $\mathbb{S}_\alpha(\mathbb{R}^d)$ of the α -stable distributions is very large with the spectral measure representation Γ . The STABLE 5.1 software simulates the general multivariate α -stable density only for dimension $d = 2$. The software, however, can simulate a special class, the sub-Gaussian α -stable density, for dimension $d \leq 100$. Section 6 contains further computational issue.

We explore subclasses of $\mathbb{S}_\alpha(\mathbb{R}^d)$, where the same reasoning applies. There is a well-known subclass called the *sub-Gaussian (or elliptically contoured)* α -stable distribution (Samorodnitsky and Taqqu, 1994; Nolan, 2013). The *isotropic* α -stable distribution is a special case of the sub-Gaussian α -stable distribution. Let us introduce the following subclasses of the α -stable distributions $\mathbb{S}_\alpha(\mathbb{R}^d)$ for $\alpha \in (0, 2]$:

- $\mathbb{IIDS}_\alpha(\mathbb{R}^d)$: the set of d independent copies of a univariate $P_{\text{stable}}(\cdot; \alpha, \sigma, \beta, \mu)$.
- $\mathbb{IDS}_\alpha(\mathbb{R}^d)$: the set of d independent univariate $P_{\text{stable}}(\cdot; \alpha_i, \sigma_i, \beta_i, \mu_i)$, $i = 1, \dots, d$.
- $\mathbb{SG}_\alpha(\mathbb{R}^d)$: the set of sub-Gaussian α -stable distributions on \mathbb{R}^d .
- $\mathbb{IS}_\alpha(\mathbb{R}^d)$: the set of isotropic α -stable distributions on \mathbb{R}^d .

Here, $\mathbb{SG}_2(\mathbb{R}^d) := \mathbb{G}(\mathbb{R}^d)$ denotes the set of Gaussian distributions. Then, it holds that $\mathbb{S}_\alpha(\mathbb{R}^d) \supset \mathbb{IDS}_\alpha(\mathbb{R}^d) \supset \mathbb{IIDS}_\alpha(\mathbb{R}^d)$ and $\mathbb{S}_\alpha(\mathbb{R}^d) \supset \mathbb{SG}_\alpha(\mathbb{R}^d) \supset \mathbb{IS}_\alpha(\mathbb{R}^d)$ for $\alpha \in (0, 2]$. Note that $\mathbb{SG}_\alpha(\mathbb{R}^d) \not\supset \mathbb{IDS}_\alpha(\mathbb{R}^d)$ and $\mathbb{IS}_\alpha(\mathbb{R}^d) \not\supset \mathbb{IIDS}_\alpha(\mathbb{R}^d)$ for $\alpha \in (0, 2)$. Gaussian distributions include independent distributions $\mathbb{G}(\mathbb{R}^d) \supset \mathbb{IDS}_2(\mathbb{R}^d)$, but this is not true for the sub-Gaussian case $\alpha \in (0, 2)$. In the following sections, we show the kernel mean of the multivariate α -stable distribution in each special case.

4.3.1 MULTIVARIATE INDEPENDENT α -STABLE DISTRIBUTION $\mathbb{IDS}_\alpha(\mathbb{R}^d)$

Let $\alpha \in (0, 2]$. Let $X = (X_1, \dots, X_d)$ be a d -dimensional independent α -stable random vector in \mathbb{R}^d , where $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$, $i = 1, \dots, d$. The kernel mean m_P of a d -dimensional independent α -stable distribution in the RKHS $\otimes_{i=1}^d \mathcal{H}_{\alpha, \tilde{\sigma}_i}$ ($\tilde{\sigma}_i > 0$, $i = 1, \dots, d$) is given by the product of the kernel means of the univariate α -stable components in (29):

$$m_P(x_1, \dots, x_d) = \prod_{i=1}^d d_{\text{stable}}(x_i; \alpha, \|(\tilde{\sigma}_i, \sigma_i)\|_\alpha, \frac{\sigma_i^\alpha \beta_i}{\|(\tilde{\sigma}_i, \sigma_i)\|_\alpha^\alpha}, \mu_i), \quad (35)$$

where $\otimes_{i=1}^d \mathcal{H}_{\alpha, \tilde{\sigma}_i}$ is the tensor product RKHS on \mathbb{R}^d generated by the p.d. kernel $\prod_{i=1}^d k_{\alpha, \tilde{\sigma}_i}(x_i, y_i)$ for $x, y \in \mathbb{R}^d$.

4.3.2 MULTIVARIATE I.I.D. α -STABLE DISTRIBUTION $\mathbb{IIDS}_\alpha(\mathbb{R}^d)$

Let $\alpha \in (0, 2]$. Let $X = (X_1, \dots, X_d)$ be a d -dimensional i.i.d. α -stable random vector drawn from $X \sim S_\alpha(\sigma, \beta, \mu)$. The kernel mean m_P of the d -dimensional i.i.d. α -stable distribution in the RKHS $\mathcal{H}_{\alpha, \tilde{\sigma}}^{\otimes d}$, ($\tilde{\sigma} > 0$) is given by

$$m_P(x_1, \dots, x_d) = \prod_{i=1}^d d_{\text{stable}}(x_i; \alpha, \|(\tilde{\sigma}, \sigma)\|_\alpha, \frac{\sigma^\alpha \beta}{\|(\tilde{\sigma}, \sigma)\|_\alpha^\alpha}, \mu), \quad (36)$$

where $\mathcal{H}_{\alpha, \tilde{\sigma}}^{\otimes d} := \otimes_{i=1}^d \mathcal{H}_{\alpha, \tilde{\sigma}}$.

4.3.3 SUB-GAUSSIAN (ELLIPTICALLY CONTOURED) α -STABLE DISTRIBUTION $\mathbb{SG}_\alpha(\mathbb{R}^d)$

Let $\alpha \in (0, 2)$ and A be a univariate $\alpha/2$ -stable random variable

$$A \sim S_{\alpha/2} \left(\left(\cos \frac{\pi \alpha}{4} \right)^{\frac{2}{\alpha}}, 1, 0 \right)$$

that is totally skewed to the right ($\beta = 1$) and $\mu = 0$. Let $G = (G_1, \dots, G_d)$ be a zero-mean Gaussian vector with a covariance matrix R in \mathbb{R}^d independent of A . Then, the random vector

$$X = A^{\frac{1}{2}} G = (A^{\frac{1}{2}} G_1, \dots, A^{\frac{1}{2}} G_d) \quad (37)$$

is called a *sub-Gaussian S α S* random vector with underlying Gaussian vector G (Samorodnitsky and Taqqu, 1994, Section 2.5, p. 77). A general sub-Gaussian α -stable random vector is a shifted sub-Gaussian *S α S* random vector with a vector $\mu^0 \in \mathbb{R}^d$. The characteristic function of the sub-Gaussian α -stable random vector is given as follows:

Proposition 4.20 (Samorodnitsky and Taqqu, 1994, Proposition 2.5.2, p. 78) *Let $\alpha \in (0, 2)$. The sub-Gaussian α -stable random vector X in \mathbb{R}^d has the characteristic function*

$$E \exp \left[i \sum_{k=1}^d \theta_k X_k \right] = \exp \left(- \left| \frac{1}{2} \sum_{ij=1}^d \theta_i \theta_j R_{ij} \right|^{\frac{\alpha}{2}} + i(\theta, \mu^0) \right), \quad (38)$$

where $R_{ij} = EG_i G_j$, $i, j = 1, \dots, d$, are covariances of the underlying Gaussian random vector $G = (G_1, \dots, G_d)$, and $\mu^0 \in \mathbb{R}^d$ is a shift vector.

Note that when $\alpha = 2$, (38) accords with the characteristic function of the Gaussian distribution (25). The sub-Gaussian is similar to the Gaussian in this sense. $\alpha = 1$ corresponds to the multivariate (non-isotropic) Cauchy distribution. Let $P_{\text{SG}}(\cdot; \alpha, R, \mu^0)$ denote the sub-Gaussian α -stable distribution with density $d_{\text{SG}}(\cdot; \alpha, R, \mu^0)$.

Corollary 4.21 *The function $k_{\alpha, R}(x, y) = d_{\text{SG}}(x - y; \alpha, R, 0)$, $x, y \in \mathbb{R}^d$, is a bounded continuous p.d. kernel and is characteristic.*

Proof Corollary 4.21 is a special case of Theorem 4.15. $d_{\text{SG}}(\cdot; \alpha, R, 0)$ with $\text{rank}(R) = d$ is a nondegenerate *S α S* density function on \mathbb{R}^d . ■

A p.d. kernel $k_{\alpha, R}(\cdot, \cdot)$ uniquely determines an RKHS $\mathcal{H}_{\alpha, R}$, and $\mathcal{H}_{\alpha, R}$ is a characteristic RKHS. The following holds for the kernel mean of the sub-Gaussian α -stable distribution $P_{\text{SG}}(\cdot; \alpha, R_1, \mu^0)$ in a sub-Gaussian RKHS $\mathcal{H}_{\alpha, R_2}$.

Proposition 4.22 *Let $\alpha \in (0, 2)$. Let $P_{\text{SG}}(\cdot; \alpha, R_1, \mu)$ be a sub-Gaussian α -stable distribution, and $\mathcal{H}_{\alpha, R_2}$ be the RKHS generated by a sub-Gaussian kernel $k_{\alpha, R_2}(x, y) = d_{\text{SG}}(x - y; \alpha, R_2, 0)$ with positive-definite matrix R_2 . If $R_1 = \sigma R_2$ for some $\sigma > 0$, the kernel mean of $P_{\text{SG}}(\cdot; \alpha, R_1, \mu)$ in $\mathcal{H}_{\alpha, R_2}$ is given by*

$$m_P = d_{\text{SG}}(\cdot; \alpha, \|(\sigma, 1)\|_{\frac{\alpha}{2}} R_2, \mu). \quad (39)$$

Proof By definition, if $R_1 = \sigma R_2$ for some $\sigma > 0$,

$$\begin{aligned} m_P &= \int_{\mathbb{R}^d} k_{\alpha, R}(x, \cdot) dP(x) = d_{\text{SG}}(\cdot; \alpha, R_2, 0) * P_{\text{stable}}(\cdot; \alpha, R_1, \mu) \\ &= d_{\text{stable}}(\cdot; \alpha, \|(\sigma, 1)\|_{\frac{\alpha}{2}} R_2, \mu). \end{aligned} \quad (40)$$

The final equality comes from the following Lemma:

Lemma 4.23 *Let $\alpha \in (0, 2)$. Let $P_{\text{SG}}(\cdot; \alpha, R_i, \mu_i^0)$, $i = 1, 2$, be two sub-Gaussian α -stable distributions. If $R_1 = \sigma R_2$ for some $\sigma > 0$,*

$$*_{i=1}^2 P_{\text{SG}}(\cdot; \alpha, R_i, \mu_i^0) = P_{\text{SG}}(\cdot; \alpha, \|(\sigma, 1)\|_{\frac{\alpha}{2}} R_2, \mu_1^0 + \mu_2^0). \quad (41)$$

Proof If $R_1 = \sigma R_2$ ($\sigma > 0$), the sum of the two log characteristic functions in (38) becomes

$$-\left|\frac{1}{2}\theta^\top R_1 \theta\right|^{\frac{\alpha}{2}} + i(\theta, \mu_1) - \left|\frac{1}{2}\theta^\top R_2 \theta\right|^{\frac{\alpha}{2}} + i(\theta, \mu_2) = -\left|\frac{1}{2}\theta^\top \|(\sigma, 1)\|_{\frac{\alpha}{2}} R_2 \theta\right|^{\frac{\alpha}{2}} + i(\theta, \mu_1 + \mu_2).$$

■

Letting $\mu_1^0 = \mu$ and $\mu_2^0 = 0$ in Lemma 4.23, we have Proposition 4.22. ■

As stated in Proposition 4.22, the kernel mean of a sub-Gaussian α -stable distribution $P_{\text{SG}}(\cdot; \alpha, R_1, \mu)$ in a sub-Gaussian RKHS $\mathcal{H}_{\alpha, R_2}$ is not always a sub-Gaussian α -stable density. If $R_1 = \sigma R_2$ with a constant $\sigma > 0$, this is true. When $\alpha \in (0, 2)$, we partition the sub-Gaussian α -stable class $\text{SG}_\alpha(\mathbb{R}^d)$ by the equivalence relation $R_1 \sim R_2$ ($R_1 = \sigma R_2$ for some constant $\sigma > 0$), so that

$$\text{SG}_\alpha(\mathbb{R}^d) = \bigcup \text{SG}_\alpha^{(d)}[R], \quad \text{SG}_\alpha^{(d)}[R] = \{P_{\text{SG}}(\cdot; \alpha, R_1, \mu) \in \text{SG}_\alpha(\mathbb{R}^d) | R_1 = \sigma R, \forall \sigma > 0\}. \quad (42)$$

The set $\text{SG}_\alpha(\mathbb{R}^d)$ of sub-Gaussian α -stable distributions is not closed under convolution, but each set $\text{SG}_\alpha^{(d)}[R]$ is closed under convolution. Note that when $\alpha = 2$, the whole set $\mathbb{G}(\mathbb{R}^d)$ of Gaussian distributions is closed under convolution. The following holds for the sub-Gaussian kernel mean w.r.t. the sub-Gaussian RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\alpha, R_0}}$.

Proposition 4.24 *Let $m_P, m_Q \in \mathcal{H}_{\alpha, R_0}$ be kernel means of sub-Gaussian distributions $P_{\text{SG}}(\cdot; \alpha, R_P, \mu_P)$ and $Q_{\text{SG}}(\cdot; \alpha, R_Q, \mu_Q)$, respectively, in the same sub-Gaussian RKHS $\mathcal{H}_{\alpha, R_0}$ generated by a sub-Gaussian kernel k_{α, R_0} . If $R_P \sim R_Q \sim R_0$, i.e., there is a constant $\sigma_P, \sigma_Q > 0$ such that $R_P = \sigma_P R_0$ and $R_Q = \sigma_Q R_0$, then*

1. $\langle m_P, k_{\alpha, R_0}(\cdot, x) \rangle_{\mathcal{H}_{\alpha, R_0}} = d_{\text{SG}}(x; \alpha, \|(\sigma_P, 1)\|_{\frac{\alpha}{2}} R_0, \mu_P), \quad x \in \mathbb{R}^d,$
2. $\langle m_P, m_Q \rangle_{\mathcal{H}_{\alpha, R_0}} = d_{\text{SG}}(0; \alpha, \|(\sigma_P, \sigma_Q, 1)\|_{\frac{\alpha}{2}} R_0, \mu_Q - \mu_P).$

Proof Statement 1 is straightforward from Proposition 2.1 and (39). Statement 2 can be obtained using Proposition 2.6, the dual of the kernel mean $\tilde{m}_P = d_{\text{SG}}(\cdot; \alpha, \|(\sigma, 1)\|_{\frac{\alpha}{2}} R_2, -\mu_P)$, and Lemma 4.23. ■

See Section 6 for an explanation of the computation of these inner products.

4.3.4 ISOTROPIC (ROTATIONALLY SYMMETRIC) α -STABLE DISTRIBUTION $\mathbb{IS}_\alpha(\mathbb{R}^d)$

If the covariance matrix R in the sub-Gaussian α -stable random vector X (Section 4.3.3) is simply $R = \sigma I$ with $\sigma > 0$ and the identity matrix I , then X is called an *isotropic (rotationally symmetric)* α -stable random vector. The set $\mathbb{IS}_\alpha(\mathbb{R}^d)$ of isotropic α -stable distributions corresponds to $\text{SG}_\alpha^{(d)}[I]$ in the partition (42). We use the notation $P_{\text{IS}}(\cdot; \alpha, \sigma, \mu^0)$, $d_{\text{IS}}(\cdot; \alpha, \sigma, \mu^0)$, $k_{\alpha, \sigma}(\cdot, \cdot)$, and $\mathcal{H}_{\alpha, \sigma}$ for the distribution, density, kernel, and RKHS, respectively. We have the following corollary for isotropic α -stable distributions.

Corollary 4.25 *Let $\alpha \in (0, 2)$. Let $P_{\text{IS}}(\cdot; \alpha, \sigma_1, \mu)$ be an isotropic α -stable distribution. Let $\mathcal{H}_{\alpha, \sigma_2}$ be the RKHS generated by an isotropic α -stable kernel $k_{\alpha, \sigma_2}(x, y) = d_{\text{IS}}(x - y; \alpha, \sigma_2, 0)$ ($\sigma_2 > 0$). The kernel mean of $P_{\text{IS}}(\cdot; \alpha, \sigma_1, \mu)$ in $\mathcal{H}_{\alpha, \sigma_2}$ is*

$$m_P = d_{\text{IS}}(\cdot; \alpha, \|(\sigma_1, \sigma_2)\|_{\frac{\alpha}{2}}, \mu) \quad (43)$$

with $\|(\sigma_1, \sigma_2)\|_{\frac{\alpha}{2}} = (\sigma_1^{\frac{\alpha}{2}} + \sigma_2^{\frac{\alpha}{2}})^{\frac{2}{\alpha}}$.

Proof Corollary 4.25 is a special case of Proposition 4.22 when $R_1 \sim R_2 \sim I$. ■

Corollary 4.26 *Let $m_P, m_Q \in \mathcal{H}_{\alpha, \sigma_0}$ be kernel means of isotropic stable distributions $P_{\text{IS}}(\cdot; \alpha, \sigma_P, \mu_P)$ and $Q_{\text{IS}}(\cdot; \alpha, \sigma_Q, \mu_Q)$, respectively, in the same RKHS $\mathcal{H}_{\alpha, \sigma_0}$ generated by an isotropic stable kernel k_{α, σ_0} . Then,*

1. $\langle m_P, k_{\alpha, \sigma_0}(\cdot, x) \rangle_{\mathcal{H}_{\alpha, \sigma_0}} = d_{\text{IS}}(x; \alpha, \|(\sigma_0, \sigma_P)\|_{\frac{\alpha}{2}}, \mu_P), \quad x \in \mathbb{R}^d,$
2. $\langle m_P, m_Q \rangle_{\mathcal{H}_{\alpha, \sigma_0}} = d_{\text{IS}}(x; \alpha, \|(\sigma_0, \sigma_P, \sigma_Q)\|_{\frac{\alpha}{2}}, \mu_Q - \mu_P).$

Proof Corollary 4.26 is a special case of Proposition 4.24 when $R_P \sim R_Q \sim R_0 \sim I$. ■

4.4 GH Kernel on \mathbb{R}^d (Multivariate GH Distribution)

The final example considers the GH distribution (Barndorff-Nielsen and Halgreen, 1977). For an overview and details, see, e.g., Prause (1999) and v. Hammerstein (2010) and references therein. Whereas the non-Gaussian stable distribution in the previous section is known to have heavy tails, the GH distribution is known to have *semi-heavy* tails (v. Hammerstein, 2010, Definition 1.12 p. 16). Hence, roughly speaking, the GH distribution lies in a class between the Gaussian distribution and the non-Gaussian stable distribution in terms of the tail property. The GH distribution is a rich model class that includes a number of probabilistic models as special cases and limiting cases. This includes NIG, hyperbolic, VG, Student-t, and Gaussian distributions. See the list in Prause (1999, Table 1.1 p.4) for contained subclasses and limiting GH distributions. The GH distribution was first applied to model the grain size distribution of windblown sand. Many applications of GH distributions, their subclasses, and related distributions are found in mathematical finance (Schoutens, 2003; Cont and Tankov, 2004; Barndorff-Nielsen and Halgreen, 1990; Madan et al., 1998; Barndorff-Nielsen, 1998; Barndorff-Nielsen and Prause, 2001; Carr et al., 2002) and utilize the Lévy process. The Matérn kernel is a well-known kernel used in spatial statistics and geostatistics. The Matérn kernel is a special case of a subclass of the SGH kernel.

Since the GH class is infinitely divisible, the SGH density, its special cases, and limiting cases¹⁰ can be used for CID kernels. The following question arises. Is the kernel mean of any GH distribution always a GH density in the RKHS associated with an SGH density?

10. If $\{P_k\}$ is a sequence of infinitely divisible distributions and $P_k \rightarrow P$ (weak convergence), then P is infinitely divisible (Sato, 1999, Lemma 7.8, p. 34).

If not, in which subclasses of GH does this hold? The first answer is no, and the second answer is the class of (zero-skewed) NIG and VG distributions.

In general, we consider the multivariate GH distribution on \mathbb{R}^d . The GH distribution is defined by a *normal mean-variance mixture* of the generalized inverse Gaussian (GIG) distribution (a definition is given in Appendix A.3). In general, the class of extended generalized Γ -convolutions is obtained by the normal mean-variance mixture of the generalized Γ -convolution (Thorin, 1978). The GH distribution is a special case of the extended generalized Γ -convolution.

Let \mathbb{P}_d denote the set of $d \times d$ positive-definite matrices. Let $\|x-y\|_\Delta := \sqrt{\langle x-y, \Delta(x-y) \rangle}$ denote the Mahalanobis distance of $x, y \in \mathbb{R}^d$ with a positive-definite matrix $\Delta \in \mathbb{P}_d$.

The multivariate GH distribution on \mathbb{R}^d has the Lebesgue density:

$$d_{GH_d(\lambda, \alpha, \beta, \delta, \mu, \Delta)}(x) = a(\lambda, \alpha, \beta, \delta, \mu, \Delta) \left(\sqrt{\delta^2 + \|x - \mu\|_{\Delta^{-1}}^2} \right)^{\lambda - \frac{d}{2}} K_{\lambda - \frac{d}{2}} \left(\alpha \sqrt{\delta^2 + \|x - \mu\|_{\Delta^{-1}}^2} \right) e^{\langle \beta, x - \mu \rangle}, \quad (44)$$

where $a(\lambda, \alpha, \beta, \delta, \mu, \Delta)$ is the normalizing constant

$$a(\lambda, \alpha, \beta, \delta, \mu, \Delta) = \frac{(\alpha^2 - \|\beta\|_\Delta^2)^{\lambda/2}}{(2\pi)^{d/2} \alpha^{\lambda-d/2} \delta^\lambda K_\lambda(\delta \sqrt{\alpha^2 - \|\beta\|_\Delta^2})}. \quad (45)$$

Here, $K_\lambda(x)$ is the modified Bessel function of the third kind with index λ . The parameter λ relates subclasses of the GH density, α is the shape parameter, β is the skewness parameter, δ gives the scaling, and μ is the location. The range of the GH parameters $(\lambda, \alpha, \beta, \delta, \mu, \Delta)$ is:

$$\lambda \in \mathbb{R}, \quad \alpha, \delta \in \mathbb{R}_+, \quad \beta, \mu \in \mathbb{R}^d, \quad \Delta \in \mathbb{P}_d, \quad \begin{array}{ll} \delta \geq 0, 0 \leq \|\beta\|_\Delta < \alpha, & \text{if } \lambda > 0, \\ \delta > 0, 0 \leq \|\beta\|_\Delta < \alpha, & \text{if } \lambda = 0, \\ \delta > 0, 0 \leq \|\beta\|_\Delta \leq \alpha, & \text{if } \lambda < 0, \end{array} \quad (46)$$

where $\delta = 0$ or $\alpha = \|\beta\|_\Delta$ correspond to limiting cases. We write $X \sim GH_d(\lambda, \alpha, \beta, \delta, \mu, \Delta)$ for the multivariate GH random vector in \mathbb{R}^d . The univariate GH distribution is given by the case $d = \Delta = 1$. The GH distribution class includes the following distributions as special cases and limiting cases:

1. $\lambda = -\frac{1}{2}$: *NIG*

$$GH_d(-\frac{1}{2}, \alpha, \beta, \delta, \mu, \Delta) = NIG_d(\alpha, \beta, \delta, \mu, \Delta). \quad (47)$$

2. $\lambda = \frac{d+1}{2}$: *hyperbolic distribution*

$$GH_d(\frac{d+1}{2}, \alpha, \beta, \delta, \mu, \Delta) = HYP_d(\alpha, \beta, \delta, \mu, \Delta). \quad (48)$$

3. $\lambda > 0$ and $\delta = 0$ (limiting case): *VG distribution*

$$GH_d(\lambda > 0, \alpha, \beta, 0, \mu, \Delta) = VG_d(\lambda, \alpha, \beta, \mu, \Delta). \quad (49)$$

4. $\lambda < 0$ and $\alpha = \|\beta\|_\Delta = 0$ (limiting case): *scaled and shifted t-distribution*

$$GH_d(\lambda < 0, 0, \mathbf{0}, \delta, \mu, \Delta) = t_d(\lambda, \delta, \mu, \Delta). \quad (50)$$

5. $\lambda < 0$ and $\alpha = \|\beta\|_\Delta \neq 0$ (limiting case): the multivariate normal mean-variance mixture with an inverse Gamma mixing distribution.

6. $\alpha, \delta \rightarrow \infty, \frac{\delta}{\alpha} \rightarrow \sigma^2 < \infty$ (limiting case): multivariate normal distribution $N_d(\mu + \sigma^2 \Delta \beta, \sigma^2 \Delta)$.

See v. Hammerstein (2010, Chapter 2) for these detailed densities. The multivariate GH distribution is symmetric if and only if $\beta = \mathbf{0}$ and $\mu = 0$. For notational simplicity, we write

1. $SGH_d(\lambda, \alpha, \delta, \Delta) = GH_d(\lambda, \alpha, 0, \delta, 0, \Delta)$,
2. $SNIG_d(\alpha, \delta, \Delta) = NIG_d(\alpha, 0, \delta, 0, \Delta)$,
3. $SHYP_d(\alpha, \delta, \Delta) = HYP_d(\alpha, 0, \delta, 0, \Delta)$,
4. $SVG_d(\lambda, \alpha, \Delta) = VG_d(\lambda, \alpha, 0, 0, \Delta)$,
5. $St_d(\lambda, \delta, \Delta) = t_d(\lambda, \delta, 0, \Delta)$.

The SGH density is the elliptical density:

$$d_{SGH_d(\lambda, \alpha, \delta, \Delta)}(x) = \frac{\alpha^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} \delta^\lambda K_\lambda(\delta \alpha)} \left(\sqrt{\delta^2 + \|x\|_{\Delta^{-1}}^2} \right)^{\lambda - \frac{d}{2}} K_{\lambda - \frac{d}{2}} \left(\alpha \sqrt{\delta^2 + \|x\|_{\Delta^{-1}}^2} \right). \quad (51)$$

The following shows that the SGH kernel on \mathbb{R}^d is characteristic. Before that, we give necessary and sufficient conditions on parameters $(\lambda, \alpha, \delta, \Delta)$ for the SGH density to be bounded.

Lemma 4.27 *The SGH density $d_{SGH_d(\lambda, \alpha, \delta, \Delta)}(x)$ on \mathbb{R}^d is unbounded if and only if we have a VG distribution, $\delta = 0$, and $0 < \lambda \leq \frac{d}{2}$, where unboundedness occurs at $x = \mathbf{0}$.*

Proof The SGH density is unbounded only if $y \downarrow 0$ in the Bessel function $K_\lambda(y)$, i.e., VG cases $\delta = 0$ and $\lambda > 0$. The VG density is unbounded if and only if $0 < \lambda \leq \frac{d}{2}$, which is straightforward from the asymptotic behavior of the Bessel function $K_\lambda(y)$ (Abramowitz and Stegun, 1968, formulas 9.6.8, 9.6.9)

$$K_\lambda(x) \sim \begin{cases} \frac{1}{2} \Gamma(|\lambda|) \left(\frac{x}{2}\right)^{-|\lambda|} & x \downarrow 0, \quad \lambda \neq 0, \\ -\ln(x) & x \downarrow 0, \quad \lambda = 0. \end{cases}$$

■

Theorem 4.28 *Let $d_{SGH_d(\lambda, \alpha, \delta, \Delta)}(x)$ be the SGH density with parameters (46). The function $k_{SGH_d(\lambda, \alpha, \delta, \Delta)}(x, y) = d_{SGH_d(\lambda, \alpha, \delta, \Delta)}(x - y)$, $x, y \in \mathbb{R}^d$, given by removing the unbounded condition in Lemma 4.27 is a continuous bounded p.d. kernel and is characteristic.*

Proof Theorem 4.28 is a special case of Theorem 3.3. ■

A p.d. kernel $k_{SGH_d(\lambda, \alpha, \delta, \Delta)}$ uniquely determines an RKHS $\mathcal{H}_{SGH_d(\lambda, \alpha, \delta, \Delta)}$, and $\mathcal{H}_{SGH_d(\lambda, \alpha, \delta, \Delta)}$ is characteristic. Since the infinitely divisible distribution class $I(\mathbb{R}^d)$ is closed under weak convergence, the subclasses and limiting cases in (47)–(50) can be used for CID kernels.

Example 4.29 *The following are CID kernels:*

1. $k_{SNIG_d(\alpha, \delta, \Delta)} = d_{SNIG_d(\alpha, \delta, \Delta)}(x - y)$, $x, y \in \mathbb{R}^d$,
2. $k_{SHYP_d(\alpha, \delta, \Delta)} = d_{SHYP_d(\alpha, \delta, \Delta)}(x - y)$, $x, y \in \mathbb{R}^d$,
3. $k_{SVG_d(\lambda, \alpha, \Delta)} = d_{SVG_d(\lambda, \alpha, \Delta)}(x - y)$, $x, y \in \mathbb{R}^d$, $\lambda > \frac{d}{2}$,
4. $k_{St_d(\lambda, \delta, \Delta)} = d_{St_d(\lambda, \delta, \Delta)}(x - y)$, $x, y \in \mathbb{R}^d$,

with associated characteristic RKHSs $\mathcal{H}_{SNIG_d(\alpha, \delta, \Delta)}$, $\mathcal{H}_{SHYP_d(\alpha, \delta, \Delta)}$, $\mathcal{H}_{SVG_d(\lambda, \alpha, \Delta)}$, and $\mathcal{H}_{St_d(\lambda, \delta, \Delta)}$, respectively.

The Matérn kernel is a special case of the SVG kernel $k_{SVG_d(\lambda, \alpha, \Delta)}$ when $\lambda > \frac{d}{2}$, $\Delta = I$, and $\alpha = \frac{\sqrt{2\nu}}{\sigma}$ (Rasmussen and Williams, 2006, Section 4.2.1) (Sriperumbudur et al., 2010, p. 1533). The following combinations allow the kernel mean to be given in a density form in the GH class.

Proposition 4.30 *When m_p is the kernel mean of density p in RKHS \mathcal{H} , consider the triplet (p, \mathcal{H}, m_p) . The following hold:*

1. $(d_{NIG_d(\alpha, 0, \delta, \mu, \Delta)}, \mathcal{H}_{SNIG_d(\alpha, \delta_0, \Delta)}, d_{NIG_d(\alpha, 0, \delta_0 + \delta, \mu, \Delta)})$,
2. $(d_{VG_d(\lambda, \alpha, 0, \mu, \Delta)}, \mathcal{H}_{SVG_d(\lambda_0, \alpha, \Delta)}, d_{VG_d(\lambda_0 + \lambda, \alpha, 0, \mu, \Delta)})$, $\lambda_0 > \frac{d}{2}$,
- 3(a). $(d_{NIG_d(\alpha, 0, \delta, \mu, \Delta)}, \mathcal{H}_{SGH_d(1/2, \alpha, \delta_0, \Delta)}, d_{GH_d(1/2, \alpha, 0, \delta_0 + \delta, \mu, \Delta)})$,
- 3(b). $(d_{GH_d(1/2, \alpha, 0, \delta, \mu, \Delta)}, \mathcal{H}_{SNIG_d(\alpha, \delta_0, \Delta)}, d_{GH_d(1/2, \alpha, 0, \delta_0 + \delta, \mu, \Delta)})$,
- 4(a). $(d_{GH_d(-\lambda, \alpha, 0, \delta, \mu, \Delta)}, \mathcal{H}_{SVG_d(\lambda, \alpha, \Delta)}, d_{GH_d(\lambda, \alpha, 0, \delta, \mu, \Delta)})$, $\lambda > \frac{d}{2}$,
- 4(b). $(d_{VG_d(\lambda, \alpha, 0, \mu, \Delta)}, \mathcal{H}_{SGH_d(-\lambda, \alpha, \delta, \Delta)}, d_{GH_d(\lambda, \alpha, 0, \delta, \mu, \Delta)})$, $\lambda > \frac{d}{2}$.

Proof of Proposition 4.30 is shown in Appendix A.4. The kernel mean of the GH distribution $P_{GH_d(\lambda, \alpha, \beta, \delta, \mu, \Delta)}$ in the RKHS $\mathcal{H}_{SGH_d(\lambda, \alpha, \delta, \Delta)}$ is not necessarily a GH density. From Proposition 4.30, the kernel mean of a zero-skewed ($\beta = 0$) NIG (resp. VG) distribution in the NIG (resp. VG) RKHS results in a zero-skewed NIG (resp. VG) density. Note that these hold only for no skewness $\beta = 0$, following Proposition A.3. Appendix A.5 shows the relevant RKHS inner product of the GH-subclass kernel mean. The computation of the inner product results in a density computation of a GH-subclass density.

5. Conjugate Kernel (Conjugate RKHS) to Probabilistic Models

Let $(I(\mathbb{R}^d), *)$ denote the convolution semigroup of infinitely divisible distributions on \mathbb{R}^d . In Section 4, we looked at several examples of kernel means of probabilistic models, where a probabilistic model, the kernel mean, and the relevant inner product have the same density form. The probabilistic models constitute sub-semigroups of $(I(\mathbb{R}^d), *)$. Based on these, we introduce the notion of the conjugate kernel (conjugate RKHS) to a set of probabilistic models in the sense of the kernel mean:

Definition 5.1 *Let $\mathcal{P}_\Theta = \{P_\theta | \theta \in \Theta\}$ be a set of probabilistic distributions P_θ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, parametrized by θ , where Θ is a subset of Euclidean space $\Theta \subseteq \mathbb{R}^d$ or an infinite dimensional set. Let \mathcal{D}_Θ be the set of probability density functions of \mathcal{P}_Θ if defined. Assume that \mathcal{D}_Θ includes a continuous, bounded positive-definite function ψ such that $\psi * P \in \mathcal{D}_\Theta$ for every $P \in \mathcal{P}_\Theta$. Then, we call the p.d. kernel $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, is conjugate to \mathcal{P}_Θ in the sense of the kernel mean. We call the associated RKHS \mathcal{H} with ψ is conjugate to \mathcal{P}_Θ .*

Suppose that $(\mathcal{P}_\Theta, *)$ is a sub-semigroup of infinitely divisible distributions $(I(\mathbb{R}^d), *)$. The following holds for the conjugate kernel to the semigroup $(\mathcal{P}_\Theta, *)$.

Proposition 5.2 *Assume that \mathcal{D}_Θ includes a symmetric continuous bounded density ψ . Then, the characteristic kernel $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, is conjugate to the semigroup $(\mathcal{P}_\Theta, *)$, i.e., $\psi * P \in \mathcal{D}_\Theta$ for every $P \in \mathcal{P}_\Theta$.*

Example 5.3 *The Gaussian kernel on \mathbb{R}^d is conjugate to the set of Gaussian distributions on \mathbb{R}^d . For each $\alpha \in (0, 2)$, the α -stable kernel on \mathbb{R}^d is conjugate to the set of α -stable distributions on \mathbb{R}^d . The sub-Gaussian α -stable kernel k_{α, R_0} on \mathbb{R}^d is conjugate to the set $\text{SG}_\alpha^{(d)}[R]$ of sub-Gaussian α -stable distributions if $R_0 \sim R$ ($R_0 = \sigma R$ for some $\sigma > 0$). The isotropic α -stable kernel on \mathbb{R}^d is conjugate to the set $\text{SG}_\alpha^{(d)}[I]$ of isotropic α -stable distributions. The NIG (resp. VG) kernel on \mathbb{R}^d is conjugate to the set of zero-skewed NIG (resp. VG) distributions. The CID kernel on \mathbb{R}^d is conjugate to the set $I(\mathbb{R}^d)$ of infinitely divisible distributions. Tempered stable distributions can also be considered as examples (Rachev et al., 2011, Table 3.2, p. 77).*

For any set \mathcal{P}_0 of probabilistic distributions, the conjugate kernel to the set \mathcal{P}_0 can be defined via the conjugate kernel to the minimal \mathcal{P}_Θ including the set \mathcal{P}_0 ($\mathcal{P}_\Theta \supseteq \mathcal{P}_0$) in the sense of Definition 5.1.

6. Computational Issue

Under the conjugate combinations, computations of kernel $k(x, y)$, kernel mean $m_P(x)$, and inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ are density computations of probabilistic models with different parameter values, such as Gaussian, α -stable, and zero-skewed NIG (VG) distributions. Note that the α -stable density in general does not have the closed-form expression except for some special cases (Appendix A.2), unlike Gaussian and NIG (VG) distributions. The α -stable density is, however, computable at least low-dimensions or special cases.

The STABLE 5.1 software provides computations of the α -stable density. For the general multivariate α -stable density $d_{\text{stable}}(x; \alpha, \Gamma, \mu)$ with the spectral measure Γ , the following

fact is exploited. Any α -stable distribution on \mathbb{R}^d can be approximated arbitrarily well by another α -stable distribution with a discrete spectral measure Γ_d , i.e., the class of α -stable distributions with discrete spectral measures Γ_d is dense in the set of all the α -stable distributions (Byczkowski et al., 1993, Theorem 1). In addition, it is known that an α -stable random vector X with a discrete spectral measure Γ_d concentrated on a finite number of points on S_{d-1} can be expressed as a linear transformation of independent α -stable random variables (Samorodnitsky and Taqqu, 1994, Proposition 2.3.7, p.70). The STABLE 5.1 software computes the α -stable density in terms of the product of the univariate α -stable densities, if the number of point masses in the discrete spectral measure Γ_d is equal to the dimension of the density. Otherwise the STABLE 5.1 computes only 2-dimensional case. The computation of the symmetric α -stable density is faster and more accurate than the nonsymmetric case. Both computations are accurate near the center of the distribution, and have limited accuracy near the tails.

For the special case of the sub-Gaussian α -stable density, the STABLE 5.1 computes the density value at $d \leq 100$. The sub-Gaussian α -stable density is elliptically contoured, and the computation is only via an amplitude density of a scalar value. The STABLE 5.1 reports that current implementation works for $\alpha \in [0.8, 2]$ and there seems to be a relative error of approximately 3% near the tails. For details, see the user manual of the STABLE 5.1. For the GH distribution, there is an R package called ghyp (Breyman and Lüthi, 2013). The zero-skewed GH density is also elliptical, and the computation reduces to evaluating the univariate amplitude density.

7. Conclusion

In this paper, we related characteristic kernels to infinitely divisible distributions in the class of nonnegative shift-invariant kernels on \mathbb{R}^d . We showed that the shift-invariant p.d. kernel is characteristic if the positive-definite function is given by a symmetric continuous bounded density of an infinitely divisible distribution. This class includes the Gaussian distribution, Laplace distribution, stable distribution, GH distribution, and other infinitely divisible distributions. The proof is immediate using the Lévy–Khintchine formula for infinitely divisible distributions. We showed how CID kernels could be generated from an (asymmetric) infinitely divisible density by the symmetrization operation. Under the guarantee that a kernel is characteristic, we gave several examples of the kernel mean of a probabilistic model, where the probabilistic model and the kernel mean have the same density form, i.e., conjugate cases. The examples ranged from the Gaussian distribution on \mathbb{R}^d to the stable distribution on \mathbb{R} , the multivariate stable distribution on \mathbb{R}^d (the sub-Gaussian and the isotropic stable classes), and subclasses of the GH distribution on \mathbb{R}^d . These are examples of sub-semigroups of the set $I(\mathbb{R}^d)$ of infinitely divisible distributions. This paper focused on proposing characteristic CID kernels, and introducing conjugate kernels for infinitely divisible distributions. The application of kernel means of (semi-)heavy tailed distributions to the density estimation, the SMM of Section 1.1, and others will be a future research. Another future direction will be the derivation of semi-parametric inference for the kernel Bayesian inference (Song et al., 2009, 2010, 2011; Fukumizu et al., 2013), using the kernel means of probabilistic models discussed in this paper.

Acknowledgments

This work was supported by JSPS KAKENHI (B) 22300098 and MEXT Grant-in-Aid for Scientific Research on Innovative Areas 25120012.

Appendix A. Supplementary Results

This section presents supplementary results.

A.1 Self-decomposable Distributions (Class L)

Let $\hat{P}(\theta)$, $\theta \in \mathbb{R}^d$, denote the characteristic function of a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$.

Definition A.1 (*Sato, 1999, Definition 15.1, p. 90*) *A probability measure $P \in \mathcal{P}(\mathbb{R}^d)$ is called self-decomposable (or of class L) if, for any $b > 1$, there is a probability measure $\rho_b \in \mathcal{P}(\mathbb{R}^d)$ such that*

$$\hat{P}(\theta) = \hat{P}(b^{-1}\theta)\hat{\rho}_b(\theta), \quad \theta \in \mathbb{R}^d. \quad (52)$$

A.2 Closed-form Solutions of Univariate α -stable Densities

In general, the α -stable density does not have a closed-form solution, except for special cases. The following are the known explicit expressions of the α -stable density. There are three special cases where univariate α -stable densities can be expressed in terms of elementary functions:

1. When $\alpha = 2$, the 2-stable distribution $S_2(\sigma, \beta, \mu)$ implies the Gaussian distribution $N(\mu, 2\sigma^2)$ with mean μ and variance $2\sigma^2$, where β has no effect, which has density

$$f_{Gauss}(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{4\sigma^2}}, x \in \mathbb{R}. \quad (53)$$

2. When $\alpha = 1$ and $\beta = 0$, the 1-stable distribution $S_1(\sigma, \beta = 0, \mu)$ implies the Cauchy distribution, which has density

$$f_{Cauchy}(x) = \frac{\sigma}{\pi((x-\mu)^2 + \sigma^2)}, x \in \mathbb{R}. \quad (54)$$

3. When $\alpha = 1/2$ and $\beta = \pm 1$, the $1/2$ -stable distribution $S_{1/2}(\sigma, \beta = 1, \mu)$ implies the Lévy distribution, which has density

$$f_{Levy}(x) = \frac{\sqrt{\sigma}}{\sqrt{2\pi}(x-\mu)^{3/2}} e^{-\frac{\sigma}{2(x-\mu)}}, \mu < x < \infty. \quad (55)$$

The kernel mean never takes the density function, since $\beta = 1$ is impossible in the kernel mean (29).

It is known that there are special cases where the univariate α -stable density can be expressed in terms of special functions. The following explicit expressions can be found in Lee (2010). Note that the α -stable kernel and the kernel mean of the α -stable distribution include some of the following explicit expressions.

Without loss of generality, we restrict ourselves to the standardized density $d_{stable}(x; \alpha, 1, \beta, 0)$ with $\sigma = 1$ and $\mu = 0$. The general density $d_{stable}(\tilde{x}; \alpha, \sigma, \beta, \mu)$ is recovered by the transformation $\tilde{x} = \sigma x + \mu$ and multiplication by $1/\sigma$.

Fresnel integrals:

When $(\alpha, \sigma, \beta, \mu) = (1/2, 1, 0, 0)$, the S α S density is

$$d_{stable}(x; 1/2, 1, 0, 0) \quad (56)$$

$$= \frac{|x|^{-\frac{3}{2}}}{\sqrt{2\pi}} \left(\sin\left(\frac{1}{4|x|}\right) \left(\frac{1}{2} - S\left(\sqrt{\frac{1}{2\pi|x|}}\right)\right) + \cos\left(\frac{1}{4|x|}\right) \left(\frac{1}{2} - C\left(\sqrt{\frac{1}{2\pi|x|}}\right)\right) \right), \quad (57)$$

where $C(z)$ and $S(z)$ are the Fresnel integrals

$$C(z) = \int_0^z \cos\left(\frac{\pi t^2}{2}\right) dt, \quad S(z) = \int_0^z \sin\left(\frac{\pi t^2}{2}\right) dt. \quad (58)$$

$k(x, y) = d_{stable}(x - y; 1/2, 1, 0, 0)$ is a CID kernel.

Modified Bessel function:

When $(\alpha, \sigma, \beta, \mu) = (1/3, 1, 1, 0)$, the one-sided continuous density is

$$d_{stable}(x; 1/3, 1, 1, 0) = \frac{1}{\pi} \frac{2^{3/2}}{3^{7/4}} x^{-3/2} K_{1/3} \left(\frac{2^{5/2}}{3^{9/4}} x^{-1/2} \right), x \geq 0, \quad (59)$$

where $K_\nu(x)$ is a modified Bessel function of the third kind. The kernel mean never takes this density function, since $\beta = 1$ is impossible in the kernel mean (29).

Hypergeometric function:

When $(\alpha, \sigma, \beta, \mu) = (4/3, 1, 0, 0)$, the S α S density is

$$\begin{aligned} d_{stable}(x; \frac{4}{3}, 1, 0, 0) &= \frac{3^{5/4} \Gamma(7/12) \Gamma(11/12)}{2^{5/2} \sqrt{\pi} \Gamma(6/12) \Gamma(8/12)} {}_2F_2 \left(\frac{7}{12}, \frac{11}{12}; \frac{6}{12}, \frac{8}{12}; \frac{3^3 x^4}{2^8} \right) \\ &\quad - \frac{3^{11/4} |x|^3 \Gamma(13/12) \Gamma(17/12)}{2^{13/2} \sqrt{\pi} \Gamma(18/12) \Gamma(15/12)} {}_2F_2 \left(\frac{13}{12}, \frac{17}{12}; \frac{18}{12}, \frac{15}{12}; \frac{3^3 x^4}{2^8} \right), x \in \mathbb{R}, \end{aligned} \quad (60)$$

where ${}_pF_q$ is the (generalized) hypergeometric function

$${}_pF_q(a_1, \dots, a_p; b_1, \dots, b_q; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n} \frac{z^n}{n!} \quad (61)$$

with Pochhammer symbol $(a)_0 = 1$, $(a)_n = a(a+1) \cdots (a+n-1)$ for $n \in \mathbb{N}^+$.

$k(x, y) = d_{stable}(x - y; \frac{4}{3}, 1, 0, 0)$ is a CID kernel.

When $(\alpha, \sigma, \beta, \mu) = (3/2, 1, 0, 0)$ (the Holtsmark distribution), the S α S density is

$$\begin{aligned} d_{stable}(x; \frac{3}{2}, 1, 0, 0) &= \frac{1}{\pi} \Gamma(5/3)_2 F_3 \left(\frac{5}{12}, \frac{11}{12}; \frac{1}{3}, \frac{1}{2}, \frac{5}{6}; -\frac{2^2 x^6}{3^6} \right) \\ &\quad - \frac{x^2}{3\pi} {}_3 F_4 \left(\frac{3}{4}, 1, \frac{5}{4}; \frac{2}{3}, \frac{5}{6}, \frac{7}{6}, \frac{4}{3}; -\frac{2^2 x^6}{3^6} \right) \\ &\quad + \frac{7x^4}{3^4 \pi} \Gamma(4/3)_2 F_3 \left(\frac{13}{12}, \frac{19}{12}; \frac{7}{6}, \frac{3}{2}, \frac{5}{3}; -\frac{2^2 x^6}{3^6} \right), x \in \mathbb{R}. \end{aligned} \quad (62)$$

The Holtsmark kernel $k(x, y) = d_{stable}(x - y; 3/2, 1, 0, 0)$ is a CID kernel.

Whittaker function:

When $(\alpha, \sigma, \beta, \mu) = (2/3, 1, 0, 0)$, the S α S density is

$$d_{stable}(x; 2/3, 1, 0, 0) = \frac{1}{2\sqrt{3\pi}|x|} \exp\left(\frac{2}{27x^2}\right) W_{-1/2, 1/6}\left(\frac{4}{27x^2}\right), x \in \mathbb{R}, \quad (63)$$

where $W_{\lambda, \mu}(z)$ is the Whittaker function defined as

$$\begin{aligned} W_{\lambda, \mu}(z) &= \frac{z^\lambda e^{-z/2}}{\Gamma(\mu - \lambda + 1/2)} \int_0^\infty e^{-t} t^{\mu - \lambda - 1/2} \left(1 + \frac{t}{z}\right)^{\mu - \lambda - 1/2} dt, \\ &\quad \text{Re}(\mu - \lambda) > -\frac{1}{2}, |\arg(z)| < \pi. \end{aligned} \quad (64)$$

$k(x, y) = d_{stable}(x - y; 2/3, 1, 0, 0)$ is a CID kernel.

When $(\alpha, \sigma, \beta, \mu) = (2/3, 1, 1, 0)$, the one-sided density is

$$d_{stable}(x; 2/3, 1, 1, 0) = \sqrt{\frac{3}{\pi}} \frac{1}{|x|} \exp\left(-\frac{16}{27x^2}\right) W_{1/2, 1/6}\left(\frac{32}{27x^2}\right), x \geq 0. \quad (65)$$

The kernel mean never takes this density function, since $\beta = 1$ is impossible in the kernel mean (29).

When $(\alpha, \sigma, \beta, \mu) = (3/2, 1, 1, 0)$, the α -stable density is

$$d_{stable}(x; 2/3, 1, 1, 0) = \begin{cases} \sqrt{\frac{3}{\pi}} \frac{1}{|x|} \exp\left(\frac{x^3}{27}\right) W_{1/2, 1/6}\left(-\frac{2}{27}x^3\right), & x < 0 \\ \frac{1}{2\sqrt{3\pi}|x|} \exp\left(\frac{x^3}{27}\right) W_{-1/2, 1/6}\left(\frac{2}{27}x^3\right), & x > 0 \end{cases} \quad (66)$$

This is the only continuous-stable case that is neither one-sided nor symmetric for which the closed-form density is available. The kernel mean never takes this density function, since $\beta = 1$ is impossible in the kernel mean (29).

Lommel function:

When $(\alpha, \sigma, \beta, \mu) = (1/3, 1, 0, 0)$, the S α S density is

$$d_{stable}(x; 1/3, 1, 0, 0) = \text{Re} \left(\frac{2 \exp(-i\pi/4)}{3\sqrt{3\pi}|x|^{3/2}} S_{0, 1/3} \left(\frac{2 \exp(i\pi/4)}{3\sqrt{3}|x|^{1/2}} \right) \right). \quad (67)$$

Here, the Lommel functions $s_{\mu,v}(z)$ and $S_{\mu,v}(z)$ are defined by

$$s_{\mu,v}(z) = \frac{\pi}{2} \left(Y_\nu(z) \int_0^z z^\mu J_\nu(z) dz - J_\nu(z) \int_0^z z^\mu Y_\nu(z) dz \right), \quad (68)$$

$$S_{\mu,v}(z) = s_{\mu,v}(z) - \frac{2^{\mu-1} \Gamma((1+\mu+\nu)/2)}{\pi \Gamma((\nu-\mu)/2)} \left(J_\nu(z) - \cos\left(\frac{\mu-\nu}{2}\pi\right) Y_\nu(z) \right), \quad (69)$$

where $J_\nu(z)$ and $Y_\nu(z)$ are Bessel functions of the first and second kind, respectively.

$k(x, y) = d_{stable}(x - y; 1/3, 1, 0, 0)$ is a CID kernel.

Landau distribution:

When $(\alpha, \sigma, \beta, \mu) = (1, 1, 1, 0)$ (the Landau distribution), the α -stable density is

$$d_{stable}(x; 1, 1, 1, 0) = \frac{1}{\pi} \int_0^\infty e^{-t \log t - xt} \sin(\pi t) dt. \quad (70)$$

The kernel mean never takes this density function, since $\beta = 1$ is impossible in the kernel mean (29).

A.3 Normal Mean Variance Mixture and GIG Distribution

The GH distribution is defined by a normal mean-variance mixture of the GIG distribution.

Definition A.2 (*v. Hammerstein, 2010, Definition 2.4, p. 78*) An \mathbb{R}^d -valued random vector X is said to have a multivariate normal mean-variance mixture distribution if

$$X \stackrel{d}{=} \mu + Z\beta + \sqrt{Z}AW, \quad (71)$$

where $\mu, \beta \in \mathbb{R}^d$, A is a \mathbb{R} -valued $d \times d$ matrix such that $\Delta := AA^\top$ is positive-definite, $W \sim N_d(0, I_d)$ is a standard normally distributed random vector, and $Z \sim G$ is an \mathbb{R}^+ -valued random variable independent of W .

A multivariate normal mean-variance mixture distribution P on \mathbb{R}^d is given by

$$P(dx) = \int_{\mathbb{R}^+} N_d(\mu + y\beta, y\Delta)(dx) G(dy), \quad (72)$$

where a probability measure G on \mathbb{R}^+ is called the *mixing distribution*. We use the simple notation $P = N_d(\mu + y\beta, y\Delta) \circ G$ to denote (72).

The GIG distribution on \mathbb{R}^+ is given by the Lebesgue density:

$$d_{GIG(\lambda, \delta, \gamma)}(x) = \left(\frac{\gamma}{\delta}\right)^\lambda \frac{1}{2K_\lambda(\delta\gamma)} x^{\lambda-1} \exp\left(-\frac{1}{2}\left(\frac{\delta^2}{x} + \gamma^2 x\right)\right) 1_{(0, \infty)}(x), \quad (73)$$

where $K_\lambda(x)$ is the modified Bessel function of the third kind with index λ . The range of the parameters $(\lambda, \delta, \gamma)$ is

$$\begin{cases} \delta \geq 0, \gamma > 0, & \text{if } \lambda > 0, \\ \delta > 0, \gamma > 0, & \text{if } \lambda = 0, \\ \delta > 0, \gamma \geq 0, & \text{if } \lambda < 0, \end{cases}$$

where $\delta = 0$ (Gamma distribution) and $\gamma = 0$ (inverse Gamma distribution) are limiting cases¹¹.

A GH distribution on \mathbb{R}^d is defined as the multivariate normal mean-variance mixture with a GIG mixing distribution:

$$GH_d(\lambda, \alpha, \beta, \delta, \mu, \Delta) := N_d(\mu + y\Delta\beta, y\Delta) \circ GIG(\lambda, \delta, \sqrt{\alpha^2 - \|\beta\|_\Delta^2}). \quad (74)$$

A.4 Proof of Proposition 4.30

Proof There is a convolution property for subclasses of the GH distribution. The following proposition is a straightforward extension of the univariate GH case (v. Hammerstein, 2010, eq. (1.9), p. 14). The proof of Proposition A.3 is given in Appendix A.6.

Proposition A.3 *For the multivariate GH class, there is a convolution stability property:*

1. $NIG_d(\alpha, \beta, \delta_1, \mu_1, \Delta) * NIG_d(\alpha, \beta, \delta_2, \mu_2, \Delta) = NIG_d(\alpha, \beta, \delta_1 + \delta_2, \mu_1 + \mu_2, \Delta),$
2. $VG_d(\lambda_1, \alpha, \beta, \mu_1, \Delta) * VG_d(\lambda_2, \alpha, \beta, \mu_2, \Delta) = VG_d(\lambda_1 + \lambda_2, \alpha, \beta, \mu_1 + \mu_2, \Delta),$
3. $NIG_d(\alpha, \beta, \delta_1, \mu_1, \Delta) * GH_d(1/2, \alpha, \beta, \delta_2, \mu_2, \Delta) = GH_d(1/2, \alpha, \beta, \delta_1 + \delta_2, \mu_1 + \mu_2, \Delta),$
4. $GH_d(-\lambda, \alpha, \beta, \delta, \mu_1, \Delta) * GH_d(\lambda, \alpha, \beta, 0, \mu_2, \Delta) = GH_d(\lambda, \alpha, \beta, \delta, \mu_1 + \mu_2, \Delta),$

where $\lambda, \lambda_1, \lambda_2 > 0$.

Statements 1 and 2 in Proposition A.3 indicate statements 1 and 2 in Proposition 4.30, letting $\beta = 0$ and $\mu_1 = 0$, respectively. Statements 3 and 4 in Proposition A.3 indicate the two statements 3(a)(b) and 4(a)(b), respectively, if we consider $NIG_d(\alpha, \beta, \delta_1, \mu_1, \Delta)$ or $GH_d(1/2, \alpha, \beta, \delta_2, \mu_2, \Delta)$ to serve as a p.d. kernel in statement 3, and $GH_d(-\lambda, \alpha, \beta, \delta, \mu_1, \Delta)$ or $GH_d(\lambda, \alpha, \beta, 0, \mu_2, \Delta)$ to serve as a p.d. kernel in statement 4. \blacksquare

A.5 RKHS Inner Product of the GH-subclass Kernel Mean

The following two statements give the RKHS inner product of the GH-subclass kernel mean.

Proposition A.4 *The following equalities hold for the multivariate GH-subclass kernel mean on \mathbb{R}^d w.r.t. the associated RKHS inner product:*

1. $\left\langle m_{P_{NIG_d(\alpha, 0, \delta, \mu, \Delta)}}, k(\cdot, x) \right\rangle_{\mathcal{H}_{SNIG_d(\alpha, \delta_0, \Delta)}} = d_{NIG_d(\alpha, 0, \delta_0 + \delta, \mu, \Delta)}(x),$
2. $\left\langle m_{P_{VG_d(\lambda, \alpha, 0, \mu, \Delta)}}, k(\cdot, x) \right\rangle_{\mathcal{H}_{SVG_d(\lambda_0, \alpha, \Delta)}} = d_{VG_d(\lambda_0 + \lambda, \alpha, 0, \mu, \Delta)}(x), \quad \lambda_0 > \frac{d}{2},$
- 3(a). $\left\langle m_{P_{NIG_d(\alpha, 0, \delta, \mu, \Delta)}}, k(\cdot, x) \right\rangle_{\mathcal{H}_{SGH_d(1/2, \alpha, \delta_0, \Delta)}} = d_{GH_d(1/2, \alpha, 0, \delta_0 + \delta, \mu, \Delta)}(x),$

11. If $\lambda \neq 0$, then $K_\lambda(x) \sim \frac{1}{2} \Gamma(|\lambda|) \left(\frac{x}{2}\right)^{-|\lambda|}$ ($x \downarrow 0$).

$$\begin{aligned}
 3(b). \quad & \left\langle m_{P_{GH_d(1/2, \alpha, 0, \delta, \mu, \Delta)}}, k(\cdot, x) \right\rangle_{\mathcal{H}_{SNIG_d(\alpha, \delta_0, \Delta)}} = d_{GH_d(1/2, \alpha, 0, \delta_0 + \delta, \mu, \Delta)}(x), \\
 4(a). \quad & \left\langle m_{P_{GH_d(-\lambda, \alpha, 0, \delta, \mu, \Delta)}}, k(\cdot, x) \right\rangle_{\mathcal{H}_{SVG_d(\lambda, \alpha, \Delta)}} = d_{GH_d(\lambda, \alpha, 0, \delta, \mu, \Delta)}(x), \quad \lambda > \frac{d}{2}, \\
 4(b). \quad & \left\langle m_{P_{VG_d(\lambda, \alpha, 0, \delta, \mu, \Delta)}}, k(\cdot, x) \right\rangle_{\mathcal{H}_{GH_d(-\lambda, \alpha, \delta, \Delta)}} = d_{GH_d(\lambda, \alpha, 0, \delta, \mu, \Delta)}(x), \quad \lambda > \frac{d}{2}.
 \end{aligned}$$

For notational simplicity, we have omitted the subscripts of kernels $k(\cdot, x)$.

Proof Proposition A.4 follows straightforwardly from Proposition 2.1 and Theorem 4.30. ■

Proposition A.5 *The following equalities hold for the multivariate GH-subclass kernel mean on \mathbb{R}^d w.r.t. the associated RKHS inner product:*

$$\begin{aligned}
 1. \quad & \left\langle m_{P_{NIG_d(\alpha, 0, \delta_P, \mu_P, \Delta)}}, m_{Q_{NIG_d(\alpha, 0, \delta_Q, \mu_Q, \Delta)}} \right\rangle_{\mathcal{H}_{SNIG_d(\alpha, \delta_0, \Delta)}} = d_{NIG_d(\alpha, 0, \delta_0 + \delta_P + \delta_Q, \mu_Q - \mu_P, \Delta)}(0), \\
 2. \quad & \left\langle m_{P_{VG_d(\lambda_P, \alpha, 0, \mu_P, \Delta)}}, m_{Q_{VG_d(\lambda_Q, \alpha, 0, \mu_Q, \Delta)}} \right\rangle_{\mathcal{H}_{SVG_d(\lambda_0, \alpha, \Delta)}} = d_{VG_d(\lambda_0 + \lambda_P + \lambda_Q, \alpha, 0, \mu_Q - \mu_P, \Delta)}(0), \\
 3(a). \quad & \left\langle m_{P_{NIG_d(\alpha, 0, \delta_P, \mu_P, \Delta)}}, m_{Q_{NIG_d(\alpha, 0, \delta_Q, \mu_Q, \Delta)}} \right\rangle_{\mathcal{H}_{SGH_d(1/2, \alpha, \delta_0, \Delta)}} = d_{GH_d(1/2, \alpha, 0, \delta_0 + \delta_P + \delta_Q, \mu_Q - \mu_P, \Delta)}(0), \\
 3(b). \quad & \left\langle m_{P_{GH_d(1/2, \alpha, 0, \delta_P, \mu_P, \Delta)}}, m_{Q_{GH_d(1/2, \alpha, 0, \delta_Q, \mu_Q, \Delta)}} \right\rangle_{\mathcal{H}_{SNIG_d(\alpha, \delta_0, \Delta)}} = d_{GH_d(1/2, \alpha, 0, \delta_0 + \delta_P + \delta_Q, \mu_Q - \mu_P, \Delta)}(0), \\
 4(a). \quad & \left\langle m_{P_{GH_d(-\lambda, \alpha, 0, \delta, \mu_P, \Delta)}}, m_{Q_{GH_d(-\lambda, \alpha, 0, \delta, \mu_Q, \Delta)}} \right\rangle_{\mathcal{H}_{SVG_d(\lambda, \alpha, \Delta)}} = d_{GH_d(\lambda, \alpha, 0, \delta, \mu_Q - \mu_P, \Delta)}(0), \\
 4(b). \quad & \left\langle m_{P_{VG_d(\lambda, \alpha, 0, \mu_P, \Delta)}}, m_{Q_{VG_d(\lambda, \alpha, 0, \mu_Q, \Delta)}} \right\rangle_{\mathcal{H}_{GH_d(-\lambda, \alpha, \delta, \Delta)}} = d_{GH_d(\lambda, \alpha, 0, \delta, \mu_Q - \mu_P, \Delta)}(0),
 \end{aligned}$$

where $\lambda_0 > \frac{d}{2}$ in statement 2 and $\lambda > \frac{d}{2}$ in statement 4(a)(b).

Proof Proposition A.5 is straightforward from Proposition 2.6 and Theorem 4.30. ■

A.6 Proof of Proposition A.3

Proposition A.3 is a straightforward extension of the convolution property of the univariate GH case (v. Hammerstein, 2010, eq. (1.9), p. 14). The normal mean-variance mixture (Definition A.2) has the following property:

Lemma A.6 *(v. Hammerstein, 2010, Lemma 2.5, p. 68) Let \mathbb{G} be a class of probability distributions on $(\mathbb{R}^+, \mathcal{B}^+)$ and $G, G_1, G_2 \in \mathbb{G}$. If $G = G_1 * G_2 \in \mathbb{G}$, then*

$$(N_d(\mu_1 + y\beta, y\Delta) \circ G_1) * (N_d(\mu_2 + y\beta, y\Delta) \circ G_2) = N_d(\mu_1 + \mu_2 + y\beta, y\Delta) \circ G. \quad (75)$$

From Lemma A.6, the normal mean-variance mixture of the convolution of two mixing distributions is the convolution of the two normal mean-variance mixtures (if $\mu = 0$, the mapping $N_d(y\beta, y\Delta) \circ G$ is a semigroup homomorphism). On the other hand, there are convolution formulas for the GIG mixing distribution:

Proposition A.7 (*v. Hammerstein, 2010, Proposition 1.11, p. 11*) *Within the class of GIG distributions, the following convolution properties hold:*

- a) $GIG(-\frac{1}{2}, \delta_1, \gamma) * GIG(-\frac{1}{2}, \delta_2, \gamma) = GIG(-\frac{1}{2}, \delta_1 + \delta_2, \gamma),$
- b) $GIG(-\frac{1}{2}, \delta_1, \gamma) * GIG(\frac{1}{2}, \delta_2, \gamma) = GIG(\frac{1}{2}, \delta_1 + \delta_2, \gamma),$
- c) $GIG(-\lambda, \delta, \gamma) * GIG(\lambda, 0, \gamma) = GIG(\lambda, \delta, \gamma), \quad \lambda > 0,$
- d) $GIG(\lambda_1, 0, \gamma) * GIG(\lambda_2, 0, \gamma) = GIG(\lambda_1 + \lambda_2, 0, \gamma), \quad \lambda_1, \lambda_2 > 0.$

The GH distribution is defined by the normal mean-variance mixture of the GIG distribution, and Lemma A.6 and Proposition A.7 prove the convolution formulas for the multivariate GH distribution (Proposition A.3).

References

- M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions (5th ed.)*. New York: Dover Publications, 1968.
- S. Amari and H. Nagaoka. *Methods of information geometry*. *American Mathematical Society*, 2000.
- E. O. Barndorff-Nielsen. Processes of normal inverse gaussian type. *Finance and Stochastics*, 2:41–68, 1998.
- E. O. Barndorff-Nielsen and K. Prause. Apparent scaling. *Finance and Stochastics*, 5: 103–113, 2001.
- O. E. Barndorff-Nielsen and C. Halgreen. Infinite divisibility of the hyperbolic and generalized inverse gaussian distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 38:309–312, 1977.
- O. E. Barndorff-Nielsen and C. Halgreen. The variance gamma (v.g.) model for share market returns. *Journal of Business*, 63:511–524, 1990.
- C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, 1984.
- M. L. Bianchi, S.T. Rachev, Y.S. Kim, and F.J. Fabozzi. Tempered infinitely divisible distributions and processes. *Theory of Probability and Its Applications (TVP), Society for Industrial and Applied Mathematics (SIAM)*, 55(1):59–86, 2010.
- B. Boots, G. Gordon, and A. Gretton. Hilbert space embeddings of predictive state representations. *UAI*, 2013.

- W. Breyermann and D. L  thi. ghyp: A package on generalized hyperbolic distributions. 2013.
- T. Byczkowski, J. P. Nolan, and B. Rajput. Approximation of multidimensional stable densities. *J. Multivar. Anal.*, 46:13–31, 1993.
- P. Carr, H. Geman, D. B. Madan, and M. Yor. The fine structure of asset returns: an empirical investigation. *Journal of Business*, 75:305–332, 2002.
- R. Cont and P. Tankov. Financial modelling with jump processes. *Boca Raton: Chapman & Hall CRC Press*, 2004.
- K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.
- K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel Dimension Reduction in Regression. *Annals of Statistics*, 37:1871–1905, 2009a.
- K. Fukumizu, A. Gretton, X. Sun, and B. Sch  lkopf. Kernel Measures of Conditional Dependence. In *NIPS*, pages 489–496. 2008.
- K. Fukumizu and C. Leng. Gradient-based kernel method for feature extraction and variable selection. In *NIPS*, pages 2123–2131. 2012.
- K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes’ Rule. In *NIPS*, pages 1737–1745, 2011.
- K. Fukumizu, L. Song, and A. Gretton. Kernel bayes’ rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, pages 3753–3783, 2013.
- K. Fukumizu, B.K. Sriperumbudur, A. Gretton, and B. Sch  lkopf. Characteristic kernels on groups and semigroups. *Advances in Neural Information Processing Systems 21*, pages 473–480, 2009b.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Sch  lkopf, and A. J. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.
- A. Gretton, K.M. Borgwardt, M. Rasch, B. Sch  lkopf, and A. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520. 2007.
- A. Gretton and L. Gy  rfi. Consistent Nonparametric Tests of Independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.
- E. Grosswald. The student t-distribution of any degree of freedom is infinitely divisible. *Zeit. Wahrsch. Verw. Gebiete*, 36:103–109, 1976.
- S. Gr  new  lder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *ICML*, pages 535–542, 2012.
- M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability. In *Proceedings of The 12th International Conference on Artificial Intelligence and Statistics*, pages 136–143, 2005.

- W. H. Lee. Continuous and discrete properties of stochastic processes. *PhD thesis, The University of Nottingham*, 2010.
- B. D. Madan, P. Carr, and E. C. Chang. The variance gamma process and option pricing. *European Finance Review*, 2:79–105, 1998.
- L. McCalman, S. O’Callaghan, and F. Ramos. Multi-modal estimation with kernel embeddings for learning motion models. In *IEEE International Conference on Robots and Automation (ICRA)*, 2013.
- K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from Distributions via Support Measure Machines. In *NIPS*, pages 10–18. 2012.
- Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert Space Embeddings of POMDPs. In *UAI*, pages 644–653, 2012.
- J. Nolan. Multivariate elliptically contoured stable distributions: theory and estimation. *Computational Statistics*, 28(5):2067–2089, 2013.
- K. Prause. *The generalized hyperbolic model: estimation, financial derivatives, and risk measures*. Ph.D. thesis University of Freiburg, 1999.
- S. T. Rachev, Y. S. Kim, M. L. Bianchi, and F. J. Fabozzi. *Financial Models with Levy Processes and Volatility Clustering*. Wiley & Sons, 2011.
- C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. *MIT Press, Cambridge, MA*, 2006.
- K. Rawlik, M. Toussaint, and S. Vijayakumar. Path Integral Control by Reproducing Kernel Hilbert Space Embedding. *Proc. 23rd Int. Joint Conference on Artificial Intelligence (IJCAI)*, 2013.
- J. Rosiński. Tempering stable processes. *Stochastic Processes and Their Applications*, 117(6):677–707, 2007.
- G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes : stochastic models with infinite variance*. Chapman & Hall, 1994.
- K. Sato. Lévy processes and infinitely divisible distributions. *Cambridge University Press*, 1999.
- B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.
- W. Schoutens. Lévy processes in finance: Pricing financial derivatives. *Chichester: Wiley*, 2003.
- D. N. Shanbhag and M. Sreehari. An extension of goldie’s result and further results in in nite divisibility. *Zeit. Wahrsch. Verw. Gebiete*, 47:19–25, 1979.
- A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

- L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert Space Embeddings of Hidden Markov Models. In *ICML*, pages 991–998, 2010.
- L. Song, K. Fukumizu, and A. Gretton. Kernel embedding of conditional distributions. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.
- L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel Belief Propagation. *Journal of Machine Learning Research - Proceedings Track*, 15:707–715, 2011.
- L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In *ICML*, pages 961–968, 2009.
- L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring Density Estimation via Reproducing Kernel Moment Matching. *ICML*, pages 992–999, 2008.
- B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.
- O. Thorin. An extension of the notion of a generalized Γ -convolution. *Scandinavian Actuarial Journal*, pages 141–149, 1978.
- E. A. F. v. Hammerstein. *Generalized hyperbolic distributions: Theory and applications to CDO pricing*. Ph.D. thesis University of Freiburg, 2010.